IMPROVEMENT OF AB INITIO METHODS OF GENE PREDICTION IN GENOMIC AND METAGENOMIC SEQUENCES

A Dissertation

Presented to

The Academic Faculty

By

Wenhan Zhu

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy in Bioinformatics

School of Biology

Georgia Institute of Technology

May, 2010

IMPROVEMENT OF AB INITIO METHODS OF

GENE PREDICTION IN GENOMIC

AND METAGENOMIC SEQUENCES

Approved by:

Dr. Mark Borodovsky, Advisor
Department of Biomedical Engineering and Computational Science and Engineering
Georgia Institute of Technology

Dr. Kostas T. Konstantinidis
School of Civil & Environmental Engineering
Georgia Institute of Technology

Dr. Jung H. Choi
School of Biology
Georgia Institute of Technology

Dr. Soojin Yi
School of Biology
Georgia Institute of Technology

Dr. King Jordan
School of Biology
Georgia Institute of Technology

Date Approved: April 1, 2010

For my Father and Mother, Zhengda and Lingdi.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Borodovsky for providing me with the opportunity to study Bioinformatics at the Georgia Institute of Technology and for his guidance, support and encouragement. I would like to thank Alex Lomsadze for his invaluable assistance, advice and friendship.

I am grateful to the members of my committee, Dr. Jung Choi, Dr. King Jordan, Dr. Kostas Konstantinidis and Dr. Soojin Yi, for their time and effort reviewing this dissertation.

This work was supported by grant HG00783 from the US National Institutes of Health (NIH), by the School of Biology and Department of Biomedical Engineering at Georgia Institute of Technology.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF SUPPLEMENTARY TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

3' UTR	3' untranslated region
5' UTR	5' untranslated region
aa	amino acid
A, T, G, C	The nucleotides of Adenine, Thymine, Guanine and Cytosine
BLAST	Basic local alignment search tool
bp	base pair
COG	Clusters of Orthologous Groups of proteins (COGs)
f	frequency
G+C %	Percent of G and C nucleotides
HMM	Hidden Markov Model
JGI	The Joint Genome Institute
kb	kilobases
KL distance	Kullback-Liebler distance
mb	megabases
NCBI	The National Center for Biotechnology Information
nr database	non-redundant database
nt	nucleotide
ORF	Open Reading Frame
P	Probability
RBS	Ribosomal Binding Site
rRNA	Ribosomal Ribonucleic Acid

Sn	Sensitivity
Sp	Specificity
tRNA	Transfer Ribonucleic Acid
URL	Uniform Resource Locator

SUMMARY

A metagenome originated from a shotgun sequencing of a microbial community is a heterogeneous mixture of rather short sequences. A vast majority of microbial species in a given community (99%) are likely to be non-cultivable. Many protein-coding regions in a new metagenome are likely to code for barely detectable homologs of already known proteins. Therefore, an *ab initio* method that would accurately identify the new genes is a vitally important tool of metagenomic sequence analysis. The standard tools for *ab initio* prokaryotic gene prediction such as EasyGene, GeneMarkS or Glimmer were not designed to work with short sequence fragments from unknown genomes. However, a heuristic model method for finding genes in short prokaryotic sequences with anonymous origin was proposed in 1999 prior to the advent of metagenomics.

The idea was to bypass traditional ways of parameter estimation such as supervised training on a set of validated genes or unsupervised training on an anonymous sequence supposed to contain a large enough number of genes. It was proposed to use dependencies between the codon frequencies and the genome nucleotide composition. In this way, the codon frequencies, critical for the model parameterization, could be derived from frequencies of nucleotides observed in the short sequence.

With hundreds of new prokaryotic genomes available it is now possible to enhance the original approach and to utilize direct polynomial and logistic approximations of oligonucleotide frequencies. This method could be further applied for initializing the

algorithms for iterative parameters estimation for prokaryotic as well as eukaryotic gene finders.

The research of this dissertation contributed to the following publications:

1) Zhu W., Lomsadze A. and Borodovsky M. (2010).

ab initio Gene Identification in Metagenomic Sequences.

Accepted, Nucleic Acids Research

2) Martin J., Zhu W., Bergman N. and Borodovsky M. (2009)

Assessment of Gene Annotation Accuracy by Inferring Transcripts from RNA-Seq.

BIBM 2009: 54-59

3) Martin J., Zhu W., Passalacqua K., Bergman N. and Borodovsky M. (2010)

Bacillus anthracis genome organization in light of whole transcriptome sequencing.

BMC Bioinformatics 2010, 11(Suppl 3):S10

4) Zhu W., Lomsadze A. and Borodovsky M.

GeneMarkS Plus: Improving gene annotation in complete prokaryotic genomes.

In Preparation.

5) Bakkeren G., Zhu W., Antonov I. and Borodovsky M.

Gene prediction in *Puccinia triticina* based on EST data.

In Preparation.

CHAPTER 1 Introduction and Literature Review

As of writing of this PhD dissertation (Spring 2010), the DNA sequences of 1,100 complete prokaryotic genomes are available to the general public through the GenBank database of the National Center for Biotechnology Information (NCBI). Since 1995, when the first bacterial genome (*Haemophilus influenzae*) (Fleischmann, Adams et al. 1995) was completely sequenced, there has been exponential growth of DNA sequencing. In about every five years, the data has been growing one more order of magnitude (Liolios, Chen et al. 2010), as depicted in Figure 1.1.

Besides the GenBank, the NCBI now provides analysis and retrieval resources of all sorts, including PubMed, Entrez, BLAST, The NCBI Taxonomy Browser, UniGene, dbSNP, dbEST and many others (Sayers, Barrett et al. 2010). Back in 2006, when I presented my PhD oral exam, I did a survey of the number of research articles returned by NCBI PubMed search using certain keywords. I did it again in February 2010, and the literature turns out to increase four to five folds, as shown in Table 1.1.

Table 1.1 The number of publications found in PubMed, by searching the keyword.
Numbers in the parenthesis correspond to the reviews.

Search keyword	May, 2006	Feb, 2010	Fold of increase
Genomics	15021 (4688)	56349 (11429)	3.8 (2.4)
Bioinformatics	12198 (1898)	69254 (11810)	5.7 (6.2)
Proteomics	7609 (2208)	23656 (5108)	3.1 (2.3)
Transcriptomics	220 (88)	791 (298)	3.6 (3.4)
System Biology	62 (18)	215 (52)	3.5 (2.9)
Interactomics	7 (5)	44 (20)	6.3 (4.0)

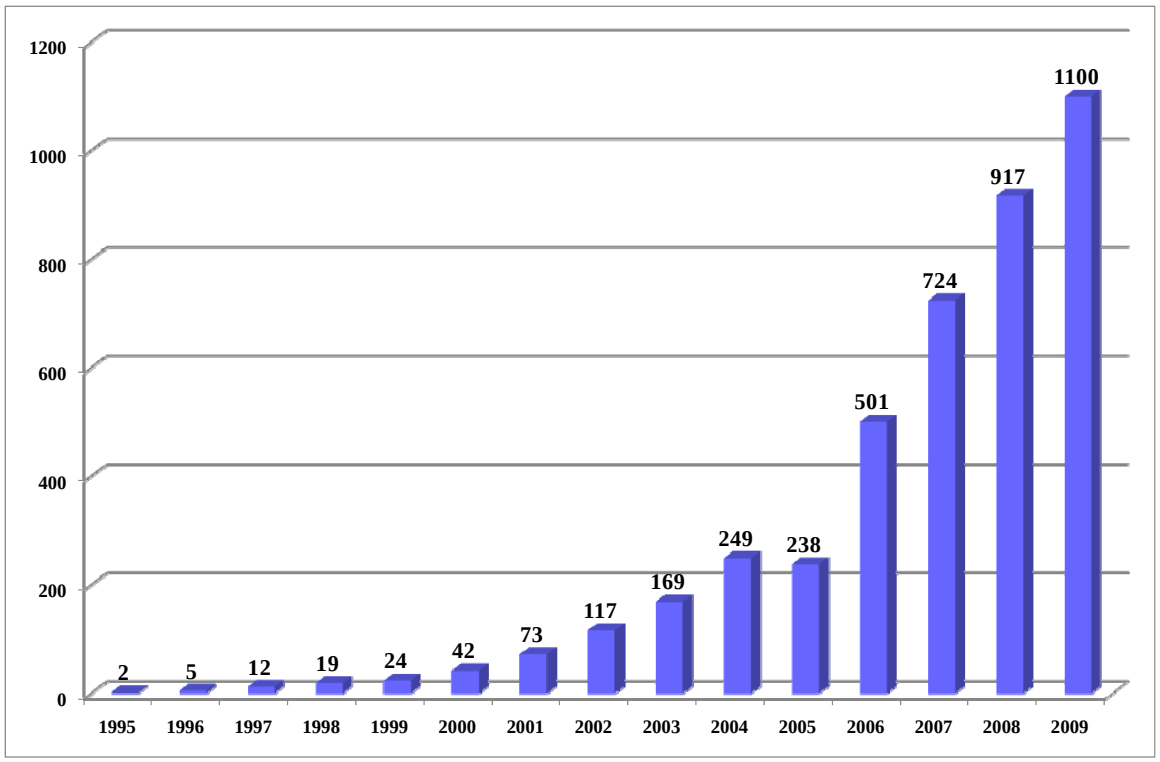


Figure 1.1 The number of completely sequenced genomes by years.
Reproduced from the data collected by The Genomes Online Database (Liolios, Chen et al. 2010) (GOLD, http://genomesonline.org/gold_statistics.htm)

Recently, there is a new type of sequence data collected from environmental samples, namely metagenomic sequences (Venter, Remington et al. 2004). These sequences usually contain many species. Low abundance samples yield small quantities of DNA that may be insufficient for library construction to sequence the genome completely (Chen and Pachter 2005). Unassembled Sanger reads and pyrosequence data could be as short as 400nt at the current level of sequencing technology.

As in the complete genome case, there are two main approaches for gene prediction. The evidence-based methods use homology searches to identify genes similar to those observed previously (Frishman, Mironov et al. 1998; Badger and Olsen 1999), commonly by BLAST similarity search (Altschul, Madden et al. 1997). Since this approach relies

entirely on comparisons to existing databases, it has major drawbacks. Low values of similarity to known sequences prevent the identification of homologs, due to either evolutionary distance or the short length of metagenomic coding sequences. Moreover, novel genes without similarities are completely ignored and would incur false negative.

On the other hand, the *ab initio* gene finding methods relies on intrinsic features of the DNA sequence to discriminate between coding and noncoding regions, allowing one to mitigate the aforementioned drawbacks (Besemer, Lomsadze et al. 2001; Lomsadze, Ter-Hovhannisyan et al. 2005; Delcher, Bratke et al. 2007; Ter-Hovhannisyan, Lomsadze et al. 2008). A heuristic model method for finding genes in short prokaryotic sequences with anonymous origin was proposed in 1999 prior to the advent of metagenomics (Besemer and Borodovsky 1999).

The focus of the work presented here is the further improvement of the heuristic method. The idea was to bypass the traditional ways of parameter estimation such as supervised training on a set of validated genes or unsupervised training on an anonymous sequence supposed to contain a large enough number of genes. It was proposed to use dependencies between the codon frequencies and the genome nucleotide composition. Therefore, the codon frequencies, critical for the model parameterization, could be derived from frequencies of nucleotides observed in the short sequence. The key observation, made upon analysis of 17 genomes, was that the frequencies of nucleotides in the three codon positions depend linearly, though with distinctly different slope coefficients, on global nucleotide frequencies (Besemer and Borodovsky 1999). In turn, due to the second Chargaff rule (Rudner, Karkas et al. 1968), this observation means that

the nucleotide frequencies in the three codon positions depend linearly on genomic GC content. These linear functions were used to reconstruct codon frequencies of the original genome from its short sequence fragment and to derive parameters of the *heuristic* second order Markov models (HAL-99 models) to be employed in a gene finding algorithm. The gene finding with heuristic models was proved to be effective for viral genomes as well as for metagenomic sequences (Mills, Rozanov et al. 2003; Kattenhorn, Mills et al. 2004).

With hundreds of new prokaryotic genomes available (Liolios, Chen et al. 2010), it is now possible to enhance the original approach and to utilize direct polynomial and logistic approximations of oligonucleotide frequencies. The analysis of the larger set of genomic sequences shows that the patterns of dependence of the codon frequencies from nucleotide frequencies are distinctly different in the two domains of life, bacteria and archaea. Interestingly, similar difference of dependences of the codon frequencies from genome nucleotide composition is observed in mesophilic and thermophilic species. Thus, for gene finding in a short sequence it is worthwhile to make a simultaneous use of two models, bacterial and archaeal, or mesophilic and thermophilic. As a by-product, the MetaGeneMark could serve as a gene domain classification program.

1.1 Dissertation outline

The chapters are organized as follows. The rest of **Chapter 1** describes a literature review of the relevant research in the area of gene finding in prokaryotic genomes and metagenomes. **Chapter 2** presents the MetaGeneMark, a novel gene finding program for metagenomics, the main topic of this dissertation. **Chapter 3** introduces and discusses the further development of GeneMarkS, a self-training algorithm for prokaryotic complete sequenced genomes.

Chapter 4 and 5 are the results from two collaboration projects. **Chapter 4** describes the application of genomic codon usage and the investigation of its relationship with gene expression level determined by RNA-Seq experiment, in the *Bacillus anthracis* genome. **Chapter 5** further applied GeneMarkS to build a refined fungus-specific gene prediction model on the *Puccinia triticina* EST data, which could be contaminated by its host genomic sequences.

1.2 Program availability

The GeneMark family programs are available for use at a website maintained by Dr. Mark Borodovsky's group at the Georgia Institute of Technology. The main URL for the GeneMark site is:

http://exon.gatech.edu/GeneMark

The MetaGeneMark application can be accessed at the URL:

http://exon.gatech.edu/GeneMark/metagenome/Prediction/

The MetaGeneMark webpage includes 1) the web interface of gene finding program for metagenomic sequences; 2) a codon usage database of 854 genomic sequences and 3) a human and mouse gut microbiome database. These results are discussed in Chapter 2 in detail.

The GeneMarkS program (Chapter 3) can be assessed at the URL:

http://exon.gatech.edu/genemarks.cgi

A GeneMarkS-predicted protein database of 313 prokaryotic genomes can be assessed at the URL:

http://exon.gatech.edu/prokaryotes_database/

The stand-alone running programs of both GeneMarkS and MetaGeneMark could be downloaded from GeneMark software distribution page:

http://exon.gatech.edu/license_download.cgi

Table 1.2 shows the usage statistics of GeneMark family programs. Other years' statistics could be found online at

http://exon.gatech.edu/stats/

Programs with * are directly related to the research of this dissertation.

Table 1.2 GeneMark programs usage per month in Year 2009

Program	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec	Subtotal
Eukaryotic GeneMark.hmm	2611	2568	4294	2698	4601	2706	7667	12922	2118	10264	3839	2049	**58337**
GeneMark 2.4 *	594	566	883	500	704	939	1133	291	540	1259	1236	1074	**9719**
GeneMarkS *	116	48	54	91	31	19	38	50	57	46	114	114	**778**
Heuristic Approach *	651	785	696	625	470	446	354	444	442	686	383	374	**6356**
Metagenome *	82	48	73	113	57	149	75	154	165	103	313	87	**1419**
Prokaryotic GeneMark.hmm *	1574	2860	2938	2111	2707	1437	1873	2662	3680	2625	2617	2160	**29244**
Download of programs	51	66	71	90	60	61	58	75	84	69	84	67	**836**
Total	**5679**	**6941**	**9009**	**6228**	**8630**	**5757**	**11198**	**16598**	**7086**	**15052**	**8586**	**5925**	**106689**

1.3 Literature review, history of gene finding in microbial genomes

1.3.1 Pioneering Finding of Statistical Patterns of Coding Sequences in Escherichia coli (1986)

Nine years before the first bacterial genomes *Haemophilus influenzae* was sequenced (Fleischmann, Adams et al. 1995), in 1986, a pioneer work was done to analyze sequenced fragments of 135,000 base pairs from the *E. coli* genome (Borodovsky, Sprizhitskii et al. 1986). This work examined the frequencies of mono and di-nucleotides and showed that the frequencies differ in the coding and non-coding DNA regions. These findings built the milestone for further development. For example, the incorporation of Bayesian approach for classification and the state-of-the-art Hidden Markov model (HMM).

The 1986 paper was done in three parts and introduced the idea progressively. The first part applied Chi-square statistical test to reject the null hypothesis that there is no dependence of two neighboring nucleotides in the coding and the non-coding regions of *E. coli* genome. In order to model this positional dependency, the uniform Markov chain alone is not sufficient. A conditional probability was introduced describing the patterns of alternation of non-uniform nucleotides, and it is the equivalent to first order Markov chain transitional probability, dependent upon previous one position of nucleotide. The second part extended to provide a second order non-uniform Markov model (formulae shown below) to determine the correlation of neighboring amino acid residues in the primary structure of *E.coli* protein molecules.

$$P(abc) = P(ab)P(c|ab), where\ a, b, c = A, T, C, G$$

This second order Markov chain requires specification of three vectors of initial probabilities of 16 components of dinucleotide frequencies, and three matrices of transitional probabilities of 4 * 16 size, in three reading frames, a total of 240 parameters ((16 + 4 * 16) * 3). A Chi-square test showed this non-uniform second order model precisely reflects the dependence of the probability of a specific triplet upon its particular positional coding frame. The authors further explored this "short-range" effect for two neighboring amino acids, i.e., two neighboring triplets / 6 nucleotides. Again, dependency was observed and the uniform Markov model could not accommodate this positional unevenness.

The last part combined the findings from the previous two steps. It analyzed three coding sequences of *E. coli* genome, ECRECA, ECLEXX and ECARAC, specified by a non-uniform first order Markov chain, a three- vector parameter set for three coding frames, while the non-coding specified by a uniform first order Markov chain, without considering the frame effect. It is significant because it applied the Bayes' formula to calculate the *posterior* probability from the a *priori* one. It analyzed only one strand of DNA sequence, but built the foundation for a remarkable work of GeneMark developed later in 1993 (Borodovsky and Mcininch 1993).

1.3.2 GeneMark (1993)

The GeneMark (Borodovsky and Mcininch 1993) improved the combination of Markov chain and Bayes approach for gene finding. Interestingly enough, Fleischmann *et. al* applied the GeneMark program to annotate protein coding genes in the first ever complete sequenced genome, *Haemophilus influenzae*. GenBank *H. influenzae* entries of 188,572 bp of protein coding sequence and 33,118 bp of noncoding sequence were used

to estimate the second order Markov chain models. The authors showed accuracies of 91.2 and 93.3 percent, in 96-bp non-overlapping coding and noncoding fragments, respectively.

Prior to 1993, the two strands of DNA sequences had to be analyzed separately, leading to false positive predictions for the strand in question while the true coding region resides on the complementary strand. Although it is possible to reduce the intensity of coding potential for a "shadow" gene using higher fifth order Markov chain (Figure 3 in (Borodovsky and Mcininch 1993)), this work used additional frame-dependent Markov chain to model the "shadow" of the coding region. It further employed the Bayes' theorem in all seven frames: the non-coding, the three coding states on the positive strand and the other three coding ones on the complementary strand.

The new GeneMark method treated the non-coding DNA sequence as a homogeneous first order Markov chain. The parameters, initial probability vector and transition matrix, were derived from the counts of mono- and dinucleotides observed from the training set of non-coding DNA sequences. The values of the transition probability matrix were approximated by dividing the dinucleotide counts by the mono ones, assuming the maximal likelihood principle. The coding regions were specified by a first-order non-homogeneous Markov chain model. It is non-homogenous in terms of the three different possible reading frames on each of the two DNA strands. The initial probability and the transition matrix were calculated in a similar fashion as the non-coding region. For a particular fragment $\boldsymbol{F}$ in question, the probability of $\boldsymbol{F}$ appearing in a coding region, for example, of frame 1, can be calculated by:

$$P(F|COD_1) = P1_0(f_1) * P1(f_2|f_1) * P2(f_3|f_2) * P3(f_4|f_3) * \ldots * P2(f_n|f_{n-1})$$

It can be repeated for the rest two coding frames on the same strand as well as the other three frames on the complementary strand (here we designate it as *Q*). Finally, the *a posteriori* probabilities that characterize the coding or non-coding property of fragment ***F*** could be determined by Bayes' theorem:

$$P(COD_m|F) = \frac{P(F|COD_m) * P(COD_m)}{\sum_j P(F|COD_j) * P(COD_j) + \sum_j Q(F|COD_j) * Q(COD_j) + P(F|NON) * P(NON)}$$

The $P(COD_m)$ and $P(NON)$ are the *a priori* probability that the starting nucleotide of fragment ***F*** falls into one out of the seven possible particular coding/non-coding frames.

The algorithm was implemented using a sliding window based approach with adjustable window and step size. For each window, a total of six *a posterior* coding probabilities were calculate and assigned to the center point of the fragment in question, and the value of one minus the total coding probability was the non-coding probability. Considering a particular open reading frame, the six coding indicator functions by the windows are averaged along the stretch of the sequence to calculate the coding potential to classify the ORF as a certain coding frame or non-coding.

The significant merit of the GeneMark algorithm was that it considered both strands of DNA simultaneously and symmetrically by introducing the shadow coding frames and incorporating it into the application of Bayes' theorem. It derived the parameter from the training set to catch its nucleotide correlations. A detailed analysis on the un-annotated regions of *E.coli* dataset EcoSeq6 was performed in 1994 (Borodovsky, Koonin et al.

1994). Results were assessed by comparing the findings of BLAST and motif search programs.

1.3.3 GeneMark-Genesis (1998)

In early 1998, Hayes and Borodovsky et al. presented the GeneMark-Genesis (Hayes and Borodovsky 1998). This was a prototype of unsupervised training procedure. It was an extension to the GeneMark program and it led to the development of GeneMark.hmm. The first new idea it introduced was namely, the root model. In the first step, all ORF's longer than a certain number (700nt) were extracted from the anonymous DNA sequence. Long ORFs of this long length are more likely to be a coding one rather than non-coding. This could be interpreted in the following way: Consider the three stop codons TAA, TAG and TGA as well as the rest 61 codons, the frequency to see a stop codon is the product of the frequencies of these three nucleotides in the particular genome. The length of ORF follows the geometric distribution with success rate equal to the sum of the three stop codons. This value depends on genomic GC-content, so that in high GC genomes, one would expect to see less stop codons, which are AT rich, leading to longer ORF's compared to low GC genomes. Claverie et al. showed that ORFs longer than 300 nt are very unlikely to occur by chance (Claverie, Poirot et al. 1997). The initial and transition probabilities of Markov models were derived upon all these long ORF sequences. This derivation didn't need the experimentally validated sets of training sequences. However, it was reliable. The model derived was called a root model. At the next step, this root model was used together with GeneMark on the anonymous sequence. This one-step derivation avoided the training and it was applicable on the novel genomes, which were just completely sequenced without genes determined by biological experiment. The

second idea worthwhile was that, the work quantitatively classified a genomic gene pool into several classes by Kullback-Leibler (KL) distance (Kullback and Leibler 1951). Each ORF could be characterized by a vector of 61 codon frequencies and it was possible to cluster ORF's into families by their vector, by calculating the KL distance of the two distinct vectors in terms of their initial and transition probabilities:

$D(P||Q) = \frac{1}{3}\sum_{k=1}^{3}\sum_{i,j=1}^{m} p_i^k p_{ij}^k log\frac{p_{ij}^k}{q_{ij}}$, where p_i^k and p_{ij}^k are the initial and transitional probabilities of the frame $\boldsymbol{k}$ in the three periodic model.

$\boldsymbol{P}$ and $\boldsymbol{Q}$ represent the coding and non-coding models $\boldsymbol{q_{ij}}$ is the transitional probabilities for the homogeneous noncoding first order Markov model and finally the sum is normalized. The geometry of the KL distance space was explored and the gene pool of the *E. coli* genes was classified into three categories: highly typical, typical and atypical.

1.3.4 GeneMark.hmm (1998)

As the algorithm name suggested, GeneMark.hmm (Lukashin and Borodovsky 1998) embedded the previously developed GeneMark approach into the framework of Hidden Markov model. The gene finding is a classification problem, i.e., to classify a certain stretch of DNA to be either coding or non-coding. In the simplest case of Hidden Markov model, every single nucleotide can be assigned one of three values 0, 1 and 2, corresponding to non-coding, coding on the direct strand and coding on the complementary strand. The core of gene finding approach is to classify the hidden states given the observed anonymous DNA sequence by HMM trajectory. To represent the prokaryotic nucleotide sequence, the GeneMark.hmm algorithm defined a total of nine hidden states, such as non-coding state, start codon state, coding state and stop codon

state. It also considered the complementary strand simultaneously by doubling the states for direct strand. Another notable aspect of the algorithm was that it used two sets of coding vectors, namely the typical and the atypical gene families, to represent the regular house-keeping and horizontally transferred genes, respectively.

The GeneMark.hmm architecture used a variant of HMM, namely, the Hidden Markov model with duration, representing the DNA as a sequence of $\boldsymbol{M}$ hidden states of a_i with duration d_i: $A = [\ (a_1 d_1)(a_2 d_2)\ \ldots\ (a_M d_M)\ ]$. Given the coding and non-coding statistics, the Viterbi algorithm (Rabiner 1989) is used to find the optimal trajectory path of hidden states A^* to go through the sequence with the maximum value of conditional probability. The probability calculation further boiled down to three parts, the probability of transition from two hidden states (in terms of gene finding, the coding and non-coding hidden states, which were unknown to classify), the probability of duration with length m, and the probability of observing the nucleotide given the hidden state. The hidden states transition probability was estimated from frequencies of native genes and foreign genes in the *E. coli* genome. The duration was derived from the length distribution of coding and non-coding regions by analytical approximation. Finally, the Markov model used in the GeneMark algorithm was readily incorporated into GeneMark.hmm, namely, a homogenous Markov model for non-coding and a three-periodic inhomogeneous Markov model for three coding frames.

The 1998 approach didn't have the hidden states to take into account of overlapping genes. But it extended the work deriving ribosomal binding site statistical models for N-terminal prediction (Hayes and Borodovsky 1998), the signal of the RBS was incorporated to fine-tune the prediction of the translation initiation site (TIS). It searched

the -19 to -4 nt upstream to the alternative start codon candidates and scored the sites by Gibbs sampling, a simulated annealing algorithm.

Later in 1999, Shmatkov *et al.* developed "frame-by-frame" algorithm (Shmatkov, Melikyan et al. 1999). As the name suggested, the HMM path decoding was parsed one frame at a time. The phases other than the frame analyzed currently were called "holes". In final step, the parses in all six reading frames were superimposed to give the prediction. This work made a trail to take into account of the overlapping genes. It defined the possible transition from the terminating stop codon to the triplet of -1, -2, …, up to –m.

1.3.5 Heuristic approach to deriving (1999)

The heuristic approach to deriving models for gene finding (Besemer and Borodovsky 1999) avoided the necessity of traditional training process, i.e., inferring the parameters of inhomogeneous Markov models for a protein coding DNA by the training sets of experimentally annotated DNA sequences. This work discovered strong linear dependencies between the positional nucleotide frequencies and the global nucleotide frequencies.

$$f(n)_i = a + b * f(n)_{global} ,$$

$$(n = A, C, G, T \ \ and \ i = position \ 1, 2 \ and \ 3 \ of \ coding \ frame)$$

Thus, the initial frequency of codon can be calculated by multiplying the three positional nucleotide frequencies, $f(abc) = f(a)_1 * f(b)_2 * f(c)_3$. The dependencies between 20 amino acid frequencies and the global genomic GC content were also found to be linear. Taking both of these into account, the authors approximated the adjusted codon

frequencies in the following way. For example, for the four-codon degenerative amino acid alanine, the frequency of one of the codons, GCT, can be calculated by this formula:

$$f_{adjusted}(GCT) = f_{alanine}\left(GC\%_{global}\right) * \frac{f_I(GCT)}{f_I(GCT) + f_I(GCC) + f_I(GCA) + f_I(GCG)}$$, here "*I*" stands for initial.

The authors found the frequencies of 10 out of the 20 amino acids changed significantly along with genome GC-content, and approximated the corresponding codon frequencies by this way. For the other 10 amino acids, the values observed in the *E. coli* (GC-content = 51%) proteome were used as constant. But still, the codon frequencies were normalized within the degenerative codon families, without the correction of these 10 amino acid frequencies.

GeneMark.hmm (Lukashin and Borodovsky 1998) had shown that a second order Hidden Markov model was sufficient for prokaryotic gene prediction due to the maximum likelihood framework to accumulate the coding signal within a long ORF. In order to reconstruct the three-periodic inhomogeneous second order model, two of the three transition probabilities, i.e., the one from the first to the second coding frame, and the one from the second to third frame, were both readily in the codon usage table. The missing information was the transition probabilities from the third to the next first coding frame, equivalent to the dependencies between two adjacent amino acids. This correlation was rather weak and it was assumed to be independent. Finally, together with the homogeneous non-coding model, all the parameters for the GeneMark.hmm program were readily available. This work was significant for the reason that it made possible to estimate the codon usage merely by a sequence as short as 400nt, while achieving

satisfactory accuracy, average of 93.1% comparing to the 93.9% by the traditional training procedure.

1.3.6 Glimmer, interpolated Markov models (1998)

Different than the Semi-Hidden Markov model or HMM with duration employed by GeneMark.hmm, Glimmer uses interpolated Markov models as its framework for capturing dependencies within oligo-nucleotides (Salzberg, Delcher et al. 1998). The authors built the training set by either extracting all ORFs longer than 500bp, or those genes with a positive BLASTp similarity found in protein database. They argued that for a fixed order Markov chain, such as the one used by GeneMark, to estimate the parameter for a k-th order, there had to be 4^{k+1} probability parameters to estimate, and all these estimates were derived from the observed occurrences of oligo-nucleotides. They proposed the IMM, to use a linear combination of probabilities dependent on the training set size, and the order can be up to eight. The score of IMM can be computed in the following way:

$$IMM_k(S_x) = \lambda_k(S_x - 1) * P_k(S_x) + [1 - \lambda_k(S_x - 1)] * IMM_{k-1}(S_x)$$

The $\lambda_k(S_x - 1)$ is the numeric weight associated with the k-mer ending at position x-1 and $P_k(S_x)$ is the estimate obtained from the training data of probability of the base located at x in the k-th order model. By induction, starting at eighth order, the IMM score can be calculated by a linear combination of the nine different order models. The authors claimed that a lower order model could be better than a higher of n-th order fixed Markov model, when there were not enough n-mers available to give a reliable estimate of the frequencies.

Glimmer 2.0 (Delcher, Harmon et al. 1999) improved over version 1.0 by introducing an interpolated context model (ICM), which was a probabilistic decision tree. The model determined the probability distribution of the base in question, conditional on a specific set context of previous bases, with the intention to use the information available maximally. The authors also made a trial to resolve the overlapping genes by scoring the overlap region and comparing the resulting neighboring two ORFs. The system attempted to move the locations of the start codons to avoid possible overlaps.

In 2007, Glimmer 3.0 included a new module to distinguish host and endo-symbiont DNA (Delcher, Bratke et al. 2007). Firstly, the authors applied the interpolated Markov model (IMM) and computed the log-odds score of any ORF in question in reverse direction, from 3' to 5'. In this way, the cumulative sum of the log-odds score would increase at the beginning of the process at 3' end, and then decline right after the presumed start codon at the 5' end, provided that the nucleotide frequency statistics of the upstream of the *true* start codon were different than those of the coding one. Similar to the application of RBS signal into gene finding by GeneMark.hmm (1998), the authors developed a post-processing program to refine the start codon positions, namely RBSfinder (Suzek, Ermolaeva et al. 2001) in 2001. The authors developed software named ELPH, similar to the Gibbs sampling, to produce the RBS positional weight matrix and score. It is integrated into Glimmer3 gene finding system. The authors concluded the work by saying Glimmer3 significantly improved over its predecessor in terms of specificity and without sacrificing the high sensitivity much.

1.3.7 Unsupervised model training program, self-training GeneMarkS (2001)

GeneMarkS (Besemer, Lomsadze et al. 2001) utilized a new iterative and non-supervised training procedure to derive the parameters for the coding and non-coding regions statistics. This unsupervised fashion of parameter estimation was introduced by (Audic and Claverie 1998), and they clustered the input anonymous DNA sequences into several partitions for machine learning of the Markov models. GeneMarkS further intergraded the previous work of GeneMark.hmm (Lukashin and Borodovsky 1998) and heuristic approach to deriving models for gene finding (Besemer and Borodovsky 1999), as well as the simulated annealing Gibbs sampling (Lawrence, Altschul et al. 1993) to localize the ribosomal binding site signal in the upstream vicinity of translation initiation site. This work didn't require the pre-defined training set of experiment verified genes for model parameters derivation. The heuristic approach discovered two linear dependencies between: i) the frequencies of a particular nucleotide in each three of its specific codon frames and its global genomic frequency; ii) the frequency of a given proteomic amino acid and the genomic GC-content. By the second Chargaff's rule (Rudner, Karkas et al. 1968), this discovery made it possible to estimate the transition and initial probability parameters from a short sample sequence, to be used for the second order homogeneous Markov model of non-coding regions as well as the three periodic inhomogeneous Markov model for coding regions.

Chapter 3 discusses more detail about further development of GeneMarkS.

1.3.8 Other HMM based methods

Several works exist with some variant of HMM architecture (Reese, Kulp et al. 2000). The GeneHacker (Yada and Hirosawa 1996) and GeneHacker Plus (Yada, Totoki et al.

2001) tried to model each single gene separately instead of the whole genome Viterbi parse. The coding model was constructed using di-codon statistics in the following way of four groups of probability values: initial probability of *P*(start codon), the second codon immediately downstream of the start codon *P*(second codon | start codon), the transition of internal codons *P*(internal | internal) and the last stop codon *P*(stop | internal). The other features were similar to GeneMark.hmm, for example, the HMM with durations and setup of RBS models.

Prior to 1994, there were several prototype of Hidden Markov model gene finding programs available (Fickett 1982; Gribskov, Devereux et al. 1984; Staden 1984; Staden 1984; Fickett and Tung 1992). Krogh et.al systematically described a complete HMM architecture, and their program was called ECOPARSE. The main HMM framework was composed of one codon HMM for the 61 triplets, its flanking stop and start codons and the intergenic region. The transition probabilities from states to states were estimated from EcoSeq6 dataset (Rudd, Miller et al. 1991). The gene model used the product of the codon probabilities as the probability of generating a stretch of coding sequences, before entering the intergenic states. This was different than the semi-Hidden Markov model (with duration) used later in GeneMark.hmm.

There were two features built in. The first was the work allowed the possibility of frame shifts to account for the sequencing errors, insertions or deletions. For each of the three nucleotides in the codon, there was a small probability, P_{indel}, set to be 10^{-8}.The second was the two types of intergenic model, one long and one short. The short model allowed 1 to 14 nucleotides while the long one was designed to capture the Shine-Delgarno sequence (Shine and Dalgarno 1974). The authors introduced the states for overlap genes,

but it didn't work well and a post processing was needed to remove the false positives. The other issue was only one strand was modeled, so that the HMM had to be applied twice on the two DNA strands, and this led to the over-prediction of "shadow genes".

This method was further improved (Krogh 1997) and a program HMMGene was developed for gene detection in Drosophila (Krogh 2000). In 2003, another major update was published. EasyGene (Larsen and Krogh 2003) was developed to estimate the statistical significance of any predicted gene. This addressed the issue that so many short ORFs were predicted as a coding gene and a significant portion of which were probably false positives. Still, HMM is used to score all the ORFs and the score was converted to a measure of significance as the expected number of ORFs that would be predicted in a random stretch of DNA sequences per mega bases. This work built the training set by extracting all ORFs longer than 120 base pairs and then using BLASTP to filter only those with significant protein match against Swiss-Port (Bairoch and Apweiler 2000) with a threshold of 10^{-5}. This procedure almost guaranteed that the training set were protein coding genes. The authors further improved the HMM by introducing a null model to model nucleotides that were not part of a gene nor in the vicinity of a gene, capturing the intergenic regions and a reverse codon model for shadow genes. As GeneMark.hmm did, the authors included RBS into the HMM. Also, three sets of codon models were used to represent the three gene classes. The other three new states are the codon positions immediately upstream of the start codon, as well as the one upstream and two codons downstream from the stop codon. The states of codon and null models are of order four and two, respectively. EasyGene used posterior decoding rather than the Viterbi algorithm for decoding. In this way, it is possible to calculate the probability that

each nucleotide was emitted by a given state *S*, by adding the probabilities that all paths having state *S* emit the nucleotide in question. Thus, the probability of a gene start would be equal to the probability of the whole gene, assuming there was no frameshift. The last step was comparing the score of all start codons in the genome. This way avoided the necessity to model the overlap genes, since each gene start would be scored independently. Finally, given the Markov statistics of the genome in question, a statistical significance measure was calculated to represent the expected number of ORFS that may be predicted with the same length-adjusted score or better in one megabase of random DNA with the same Markov properties. As an application, the authors applied EasyGene on a total of 143 prokaryotic genomes (Nielsen and Krogh 2005) and found that some of the genomes were over-annotated in RefSeq database, due to many short ORFs were annotated as protein coding genes, this was a further extension of their work in 2001 (Skovgaard, Jensen et al. 2001).

1.3.9 Similarity search based gene prediction methods

Other than the *ab initio* gene finding methods, there were several methods used similarity search based on the encoded protein. Such extrinsic analysis involves BLAST (Altschul, Madden et al. 1997) type mapping of candidate gene products against protein sequence databanks. ORPHEUS (Frishman, Mironov et al. 1998) was the first such program, and it extracted most possible reliable ORFs as the 'seed ORF' by BLAST search, then used this set to estimate coding potential parameters. CRITICA is another one, and its name stands for coding region identification tool invoking comparative analysis (Badger and Olsen 1999). While ORPHEUS extracted the training sequences from the PIR-International protein sequence databank, CRITICA directly applied the BLASTN to find

similarity on the DNA level. A di-codon usage then was estimated and applied into the gene finding in *Salmonella typhimurium* genome.

1.3.10 Other methods, Support Vector Machine (SVM) and Neural Networks

Recently, Krause *et. al* described a novel gene finder GISMO (gene identification using a support vector machine for ORF classification) (Krause, McHardy et al. 2007). Support vector machine (SVM) attracted substantial research interest (Cristianini and Shawe-Taylor 2000). As a maximum margin classifier, SVM learns an optimally separating hyper plane in a higher dimension feature space. To apply SVM into gene finding, the authors used domain motif as the classification target for the CDS, instead of score the complete ORF. The parameter of the other class, the nORF as called by them, were learned from the shadow of the coding genes, namely the statistics from the other reading frames. Ten features were evaluated, and they were oligo-nucleotides of length 3-9, mono- and di-amino acids, as well as a combination of codons and acids. Their result showed that the 64 dimensional vectors of relative in-frame codon frequencies performed the best, and this finding was in agreement of traditional HMM-based gene finding methods: the major dependencies of neighboring coding nucleotides were of second order Markov chain. The other distinction was that the authors built the training set using those proteins with domain similarity to Pfam-A database (Finn, Tate et al. 2008).

The GeneMark-Genesis program (Hayes and Borodovsky 1998) derived two models for each genome according to typical and atypical codon usage clusters. The RescueNet (Mahony, McInerney et al. 2004) revisited this issue of intra-genomic compositional variation using Self-Organizing Map (SOM), amongst several neural network based gene finding methods (Xu, Mural et al. 1994; Xu and Uberbacher 1996). The authors used the

relative synonymous codon usage (RCSU) as the measure of gene coding potential. RCSU is the quotient of dividing the observed number of occurrences of a particular codon by the expectation. Trp, Met and three stop codons were not accounted and thus a total of 59-number vector was used in the SOM.

1.3.11 Frameshift detection programs

The other potential problem to gene finding is the frame shift, either programmed or due to sequencing errors. Posfai *et al.* developed an extrinsic algorithm based on protein similarity search (Posfai and Roberts 1992). The idea is that an insertion or deletion error within a coding sequence would interrupt the reading frame. Such errors can be detected by comparing any known protein sequence in database against the conceptual translation protein product of the DNA sequences in all six reading frames. This approach was further extended by scanning the query nucleotide sequence against databases of protein sequences and effectively hybridizing similar fragments onto the query in any of its six reading frames (Brown, Sander et al. 1998). On the other hand, the intrinsic (*ab initio*) approach tries to give the solution without requiring the subject protein database. GeneMark has a subroutine to identify possible frameshifts (Borodovsky and Mcininch 1993). FrameD was initially designed as gene prediction program with focus on finding possible frameshifts (Schiex, Gouzy et al. 2003). It uses a weighted directed acyclic graph, with seven tracks for the six reading frames and one non-coding frame. Every path in the graph represents a possible gene prediction. Edges between two coding tracks are graphical representation for potential frame shift; while the deletion was modeled as the edge jumping over one nucleotide. The probability associated with the edges is defined by the emission probability of a track-specific interpolated Markov Model (0^{th} order

Markov model for the non-coding and three-periodic IMM for the coding tracks). FrameD computes for every edge in the graph the probability that it is used over all possible predictions to form the optimal path, using a forward-backward like dynamic programming algorithm. Recently, GeneMark was further extended to process the posterior probabilities of hidden states to detect the *jumps* between the coding frames (Kislyuk, Lomsadze et al. 2009).

1.3.12 Gene finding in metagenomic sequences

Modified versions of Fgenesb, GeneMark.hmm and Glimmer were used to predict genes on the metagenomic datasets by different sequencing centers (Lukashin and Borodovsky 1998; Delcher, Bratke et al. 2007). In 2006, MetaGene (Noguchi, Park et al. 2006) further extended the non-supervised training procedures (Besemer and Borodovsky 1999) and used logistic regression to estimate the codon frequency by G+C content. The authors confirmed the strong dependencies between mono-/di-codon frequencies and the GC% of genomic sequence in bacteria and archaea, corresponding to two sets of regression formulas. For any anonymous input sequence, a domain classification step is first applied to score the open reading frames, and then a higher optimal path is selected to predict the kingdom and protein-coding genes. Other features implemented include the consideration of length distribution of ORFs, distance distribution from the correct start codon to the leftmost start codon and the orientation and distance of neighboring ORFs. Later in 2008, an improved version, MetaGeneAnnotator (Noguchi, Taniguchi et al. 2008) used two additional models, a self-training model for long sequences such as complete genomes and a model to detect prophage genes in addition to the chromosome genes. The self-training model is derived by weighting the average of the di-codon frequencies from the

predicted genes and from the regression models used for the initial prediction. As the RBS was used in GeneMarkS (Besemer, Lomsadze et al. 2001), the authors analyzed the upstream sequences of annotated genes from 229 prokaryotic genomes. They derived nine fixed motifs to represent the species-specific pattern of RBSs. For very short sequences (having no training data), a general model of the RBS was constructed based on average RBS map. Accuracy for 700nt fragments reaches the level of 96% sensitivity and 93% specificity.

Also, neural network was also used to classify ORF in metagenomic sequences. The program Orphelia (Hoff, Tech et al. 2008) uses seven sequence characteristics features, which include mono- and di-codon usage, TIS coverage and probability, length scores of complete and incomplete genes, as well as the GC-content. Their web server paper (Hoff, Lingner et al. 2009) compared the gene prediction in the JGI-FAMeS dataset (Mavromatis, Ivanova et al. 2007), against MetaGene and FGENESB.

Regarding the gene 5' start codon prediction in metagenomic sequences, MetaTISA (Hu, Guo et al. 2009) added one more step of binning procedure into the scheme that was used in TriTISA. The binning procedure employs a Bayesian classifier based on *k*-mer frequencies. Fragments from the same phylogenetic groups are assumed to have close origin and share the translation initiation mechanism. Three posterior probabilities are calculated for each candidate TIS, and they are 1) the candidate as a true TIS 2) the candidate from non-coding region and 3) the candidate from coding region. They compared to the current version of MetaGeneAnnotator (Noguchi, Taniguchi et al. 2008) and showed improved accuracy in a test set of six genomes.

1.3.13 Motif finding – Several types of Gibbs Sampler

Gibbs sampler is a major component used by GeneMarkS to locate the signal ribosomal binding site in the prestart regions of microbial genomes. The 2001 GeneMarkS version used Gibbs site sampler version 1.0, a detail description of the algorithm could be found in (Lawrence, Altschul et al. 1993).

1.3.13.1 Gibbs Site Sampler (1993)

Pattern recognition in multiple proteins or nucleic acids sequences could lead to important discovery of functional motifs. The two variants, global and local multiple alignment, were especially of interest. Gibbs sampler was developed in 1993 to solve the local multiple alignment, assuming no prior information on the patterns and locations within the sequences (Lawrence, Altschul et al. 1993). It was a heuristic Markov Chain Monte Carlo (MCMC) algorithm and has been applied in the GeneMark.hmm algorithm to locate the ribosomal binding site upstream to the gene start 5' regions.

The algorithm tried to solve the problem of finding a relatively small number of sequence patterns, consisting of one un-gapped segment from each of the input sequences. This pattern can be modeled by two parts: a probabilistic model of residue frequencies at each position and a location pattern described by a set of probabilistically inferred position patterns. The optimization procedure applied a stochastic expectation maximization (EM) method by iterative sampling. It is initialized by choosing random starting positions within the various sequences and then proceeds through many iterations of predictive update step in the following fashion: In each iteration, one of the sequences was chosen randomly. The motif description and the background frequencies were updated to exclude this sequence. Secondly, each possible position within the sequence was sampled to

calculate the probability of generating the expected motif according to the current pattern probabilities as well as the background probabilities. The ratio of these two was used as the weight to be assigned to that position in question. And finally, a stochastic position was selected based on the weights and the sequence was put back into the sequence pool to update both the motif and the background models. The core of the algorithms lies on that, once some correct motif positions were selected randomly, the motif probability matrix would begin to reflect the true pattern within other sequences and finally converge.

One defect of the algorithm was that it could get stuck into the local maximum without reaching the global maximum. The authors suggested two ways to resolve: phase shift and more number of iterations. The phase shift altered motif position by a certain number and compared the probability ratios to make a selection stochastically. They tried out a test set of 30 helix-turn-helix motif sequences, and it was shown the convergence was achieved after 4,000 sampling iterations.

1.3.13.2 Gibbs Motif Sampler (1995)

Later in 1995, the authors updated the algorithm, called motif sampling (Neuwald, Liu et al. 1995), to address the problem of detecting motifs with little prior information available, while the earlier version was called site sampler. The site sampler assumed a fixed positive number of motifs within each sequence and then iteratively sampled the sites. The motif sampler partitioned the input sequences into regions corresponding to a specified number of models (including a null model representing no motifs at all). An alignment of the sequence was constructed separately as several segments, each with a corresponding residue frequency model. In the sampling step, a site was randomly selected in a similar way as the site sampler. The difference was that, several values of

likelihood were calculated for each of the models, as well as the possible null model. In this way, each model was weighted by the posterior probability that an arbitrary site belonged to that model in question. Finally, an update model was selected stochastically based on the weight.

1.3.13.3 Gibbs Recursive Sampler (2003)

A variation of the Gibbs Motif Sampler, the Gibbs Recursive Sampler (Thompson, Rouchka et al. 2003) was developed by specifically for locating multiple transcription factor binding sites (TFBS) in unaligned and heterogeneous DNA sequences. It was designed to search for multiple TFBS simultaneously using a rigorous Bayesian method for inferring the number and the locations of the TFBS for multiple TF motifs simultaneously.

Multiple transcription factors often bind in a combinatorial fashion to regulate transcription, so that the exact number of sites and the number of sites corresponding to each motif in any input sequence are unknown and often vary. The recursion algorithm examines the placements of sites and infers for each sequence the total number of sites, the number of each of the motifs and then the alignments. The sampling process iterates over the sequences one at a time, calculates the score based on the current alignment and then guides the sampling process toward convergence. After the sampling, with a set of multiple motif position determined, the log of the posterior alignment probability is calculated. The maximum *a posterior* probability (MAP) is measured relative to a null alignment, by taking the difference between the log of the probability of the alignment and the log of the probability of an empty alignment, a greater value than zero indicates a more likelihood of the alignment than the unaligned background.

There were two major improvements over previous versions, using heterogeneous background and prior information of binding motifs. Earlier development assumed homogeneous background models in the composition of each sequences. A background model to take into account of the heterogeneity in the composition of background nucleotide sequence was implemented. The Bayesian segmentation algorithm (Liu and Lawrence 1999) calculates the probabilities of observing each of the four bases at each position in a sequence, namely, a positional specific frequency matrix, for the log likelihood calculation of the background null model. In some datasets, a sufficient number of sites from prior studies could be available. This recursive sampler version can convert this piece of information to a prior position weight matrix motif model. This informed prior model provides clues to the expected pattern in DNA binding motifs but fortunately, it does not affect the posterior inference of sites and motifs. Based on this work, the authors analyzed two data sets of clusters of TFBSs (Thompson, Palumbo et al. 2004). The algorithm finds 69% of experimentally reported TFBSs in one set and 85% of the *cis*-regulatory modules in the other reference data set of regions upstream of genes differentially expressed in skeletal muscle cells.

The latest version, the Centroid sampler, improved by minimizing the pairwise distance between sampling runs. The Centroid version was used with phylogenetic tree to analyze promoter data from closely related species (Newberg, Thompson et al. 2007). This version is not suitable for gene finding in microbial genomes.

CHAPTER 2 *ab initio* Gene Identification in Metagenomic Sequences

Abstract

We describe an algorithm for gene identification in DNA sequences derived from shotgun sequencing of microbial communities. Accurate *ab initio* gene prediction in a short nucleotide sequence of anonymous origin is hampered by uncertainty in model parameters. While several machine learning approaches could be proposed to bypass this difficulty one effective method is to estimate parameters from dependencies, formed in evolution, between frequencies of oligonucleotide in protein-coding regions and genome nucleotide composition. Original version of the method was proposed in 1999 and has been used since for i/ reconstructing codon frequency vector needed for gene finding in viral genomes and ii/ initializing parameters of self-training gene finding algorithms. With advent of new prokaryotic genomes en masse it became possible to enhance the original approach by using direct polynomial and logistic approximations of oligonucleotide frequencies as well as by separating models for bacteria and archaea. These advances have increased the accuracy of models reconstruction and, subsequently, gene prediction. We describe the refined method and assess its accuracy on known prokaryotic genomes split into short sequences. Also, we show that as a result of application of the new method, several thousands of new genes could be added to existing annotations of several human and mouse gut metagenomes.

2.1 Introduction

A metagenomic sample is a heterogeneous mixture of rather short sequences originated from a shotgun sequencing of a microbial community. A vast majority of microbial species in a given community (99%) are likely to be non-cultivable (Chen and Pachter 2005). Many protein-coding regions in a new metagenome are likely to code for barely detectable homologs of already known proteins. Therefore, along with comparative genomic methods that relay on sequence similarity search, *ab initio* methods able to identify genes having no similarity to ones existing in databases are vitally important tools of metagenomic sequence analysis. Sequence similarity based methods possess high specificity and ability to characterize function of predicted genes (Venter, Remington et al. 2004; Krause, Diaz et al. 2006; Yooseph, Sutton et al. 2007; Yooseph, Li et al. 2008). *Ab initio* gene finders exhibit high sensitivity along with sufficiently high specificity. The standard tools for *ab initio* prokaryotic gene prediction such as EasyGene (Larsen and Krogh 2003), GeneMarkS (Besemer, Lomsadze et al. 2001) or Glimmer (Delcher, Bratke et al. 2007) were not designed to work with short sequence fragments from unknown genomes. However, a special method for assignment of parameters of GeneMark.hmm, the *heuristic model* method, designed for accurate gene finding in short prokaryotic sequences with anonymous origin was proposed four years prior to the advent of metagenomics (Besemer and Borodovsky 1999).

The idea was to bypass traditional ways of parameter estimation such as supervised training on a set of validated genes or unsupervised training on an anonymous sequence supposed to contain a large enough number of genes. It was proposed to use dependencies, apparently formed in evolution, between codon frequencies and genome

nucleotide composition. Therefore, the vector of codon frequencies, critical for the model parameterization, could be derived from frequencies of nucleotides observed in a short sequence. This *heuristic model* method has been used for i/ reconstructing codon frequency vector for gene finding in viral genomes (Mills, Rozanov et al. 2003) and ii/ initializing the algorithms for iterative parameters estimation for prokaryotic as well as eukaryotic gene finders (Besemer, Lomsadze et al. 2001; Lomsadze, Ter-Hovhannisyan et al. 2005; Ter-Hovhannisyan, Lomsadze et al. 2008). Recently, several new methods for *ab initio* gene finding in metagenomic sequences have been developed (Noguchi, Park et al. 2006; Hoff, Tech et al. 2008; Noguchi, Taniguchi et al. 2008). Particularly, the authors of MetaGene (Noguchi, Park et al. 2006) saw a significant potential in the *heuristic model* method (Besemer 1999); they have extended the method to use of di-codon frequencies. The authors of the new tools have shown that their performance is comparable to performance of the original *heuristic model* method (Noguchi, Park et al. 2006; Suppl. Table 3) (Noguchi, Park et al. 2006; Hoff, Lingner et al. 2009).

In this paper we describe further improvement of the *heuristic model* method. A key observation made upon analysis of 17 genomes (Besemer and Borodovsky 1999) was that frequencies of nucleotides in the three codon positions depend linearly, though with distinctly different slope coefficients, on global nucleotide frequencies. In turn, due to the second Chargaff rule (Rudner, Karkas et al. 1968), this observation means that nucleotide frequencies in the three codon positions depend linearly on genomic GC content. These linear functions were used to reconstruct codon frequencies of the original genome using information derived from its short sequence fragment and to derive parameters of the *heuristic* second order Markov models (HAL-99 models) for a gene finding algorithm.

Gene finding with *heuristic* models was proved to be effective for viral genomes (Mills, Rozanov et al. 2003; Kattenhorn, Mills et al. 2004) as well as for metagenomic sequences.

With hundreds of new prokaryotic genomes available it is now possible to enhance the original approach and to utilize direct polynomial and logistic approximations of oligonucleotide frequencies. Also, analysis of a larger set of genomic sequences has shown that patterns of dependence of codon frequencies from nucleotide frequencies are distinctly different in the two domains of life, bacteria and archaea. Interestingly, distinctly different patterns of dependence of codon frequencies from genome nucleotide composition have also been observed in mesophilic and thermophilic species. Thus, for gene finding in a short sequence it is worthwhile to make a simultaneous use of two models, bacterial and archaeal, or mesophilic and thermophilic.

We have assessed an accuracy of a gene finder, GeneMark.hmm, using the new models on the sets of short sequences obtained by splitting known genomes into equal length fragments (in a range from 72nt to 1100nt). The results demonstrate a higher accuracy in comparison with several other existing methods as well as with the use of original heuristic models.

Application of whole-genome shotgun sequencing to studies of mixed microbial communities, such as gut microbiota of human and mouse has a potential to reveal details of a large picture of the host metabolism combining microbial and mammalian elements. It is estimated that human intestinal microbiota consists of 10^{13} to 10^{14} microorganisms. This microbiome should contain at least 100 times as many genes as human genome *per se*. Still, due to diversity of the microbiome, metagenomic datasets consist mainly of

unassembled single-read sequences. We have applied the new method to the sequences of human and mouse gut microbial communities (Gill, Pop et al. 2006; Turnbaugh 2006). We detected a large number of protein-coding regions not yet annotated; for a significant fraction of the protein products of newly predicted genes we found homologues among known proteins. Notably, identification of incomplete genes carries valuable information for reconstruction of metabolic networks and signaling pathways. Since a number of protein-coding regions in a metagenome may be counted by millions (Yooseph, Sutton et al. 2007), improving accuracy of gene finding by a percentage point would affect accurate prediction of tens of thousands of genes of the organisms constituting microbial communities. Therefore, development of accurate metagenome specific methods is of critical importance for quality analysis of sequence data produced by the next generation sequencing technologies (Mardis 2008).

2.2 Materials

2.2.1 Training set

Sequence data of 582 complete prokaryotic genomes (534 bacteria and 48 archaea; genetic code 11) were from the NCBI RefSeq database. A length of the shortest genome in the sample, *Nanoarchaeum equitans* (Randau, Munch et al. 2005), was 490 Kbp. The genome GC content varied from 16.6% to 74.9%. The data on optimal growth temperature for 357 prokaryotic species (Supplementary Table 1) was from the NCBI Entrez genome database (Sayers, Barrett et al. 2009). Metagenomic sequence data and annotation for human and mouse gut microbiomes were from the JGI IMG/M database (Markowitz, Ivanova et al. 2008).

2.2.2 Test set

For assessment of gene prediction accuracy we used fragments from whole genomes of 29 bacterial and 15 archaeal species listed in Supplementary Table 2. The genomes were split into equal length non-overlapping fragments, with length in range from 72nt to 1100nt; fragment annotations were derived from corresponding RefSeq records. To retain genes with most reliable annotation, fragments overlapping annotated hypothetical genes were discarded.

2.3 Methods

2.3.1 Heuristic method of model parameters derivation

A conventional *ab initio* gene finding algorithm employs a probabilistic model of genomic sequence containing protein-coding and non-coding regions. Gene prediction accuracy critically depends on precision of estimation of model parameters which are genome specific. The number of parameters of probabilistic model of a protein-coding region, a three-periodic Markov chain model (Borodovsky and Mcininch 1993) increases exponentially ($\sim 4^N$) with the Markov chain order N. The higher is the model order the larger is the size of a set of training sequences required for parameter estimation without over-fitting, e.g., in practice, estimation of parameters of the fifth order model is made on a set of verified protein-coding sequences with total length of 400,000nt. Note that in our observations even if a larger training set is available, models with an order higher than five did not make a noticeable difference in power of discrimination between coding and non-coding regions (Azad and Borodovsky 2004).

Metagenomic sequence data, mixtures of shotgun sequences from numerous members of microbial communities, are populated with short sequences (with length as short as 400nt and even shorter). The task is to identify a complete or incomplete protein-coding region residing in a short fragment. A gene finding algorithm, e.g. GeneMark.hmm, could be applied to solve this task should we know or are able to derive the genome specific model parameters. However, the fact that genomic context of the short fragment is missing precludes a use of standard approaches for parameter estimation. In a previous work (Besemer and Borodovsky 1999) we proposed a method to infer parameters of the three-periodic second order Markov model for gene finding in a short (e.g. 400nt) sequence fragment of unknown origin. First, we have identified dependencies that link the nucleotide composition of a genome with the genome specific codon frequencies. These dependencies are apparently the strongest factors that determine a genome wide synonymous codon usage pattern (Knight, Freeland et al. 2001; Chen, Lee et al. 2004). Second, nucleotide frequencies observed in the short DNA fragment served as estimates of global nucleotide frequencies in the whole genome, the source of short fragment. Then, starting from estimated values of global nucleotide frequencies we reconstructed the genome specific codon frequencies.

In more details, in the first step, analysis of genomes with known annotation, by taking one genome at a time we determined frequencies of occurrence of each of 61 codons in a genome-wide set of annotated protein-coding regions. The codon frequency data determines 12 genome specific positional frequencies f_{1X}, f_{2X}, and f_{3X} X = A,C,G,T, in the three codon positions. For a sample of known genomes $r=1,2,\ \ldots\ R$ with observed f_{kX}, k = 1,2,3 the f_{kX} values were approximated by linear regression on the global nucleotide

frequency f_X, X = A,C,G,T. Initially, in 1999, the analysis was done for 17 completely sequenced genomes (Fig. 1 in (Besemer and Borodovsky 1999), see also (Gorban and Zinovyev 2007)).

Now, with many more sequenced genomes available the linear regression analysis was done for 319 bacterial genomes (Figure 2.1) as well as for 38 archaeal genomes (Table 2.1a). Graphs in Figure 2.1 look different from graphs in Fig. 1 in (Besemer and Borodovsky 1999) for the following reasons. The global nucleotide frequency variable strongly correlates with the genome GC content. The second Chargaff rule states that at a whole genome level, nucleotide frequencies, f_X, X = A,C,G,T in a single DNA strand are such that $f_A \sim f_T$ and $f_G \sim f_C$. Therefore, four nucleotide frequencies observed in whole genomes can be derived from a single parameter, the GC content; if s is a genomic GC content, $f_G + f_C$, then frequencies of nucleotides $f_G = f_C = s / 2$ and $f_A = f_T = (1\text{-}s) / 2$. Thus, new graphs of positional nucleotide frequencies (Figure 2.1) were plotted as functions of genomic GC content.

Further, the s value determined for a short genomic fragment is used as predictor of positional nucleotide frequencies f_{kX}, k = 1,2,3 and X = A,C,G,T. Assuming that a codon frequency, f_{XYZ}, is proportional to product $f_{1X}f_{2Y}f_{3Z}$ we could obtain an initial approximation of codon frequency f'_{XYZ}. Additional correction comes from the value of predicted frequency of encoded amino acid α, f_α(s) determined by linear regression of

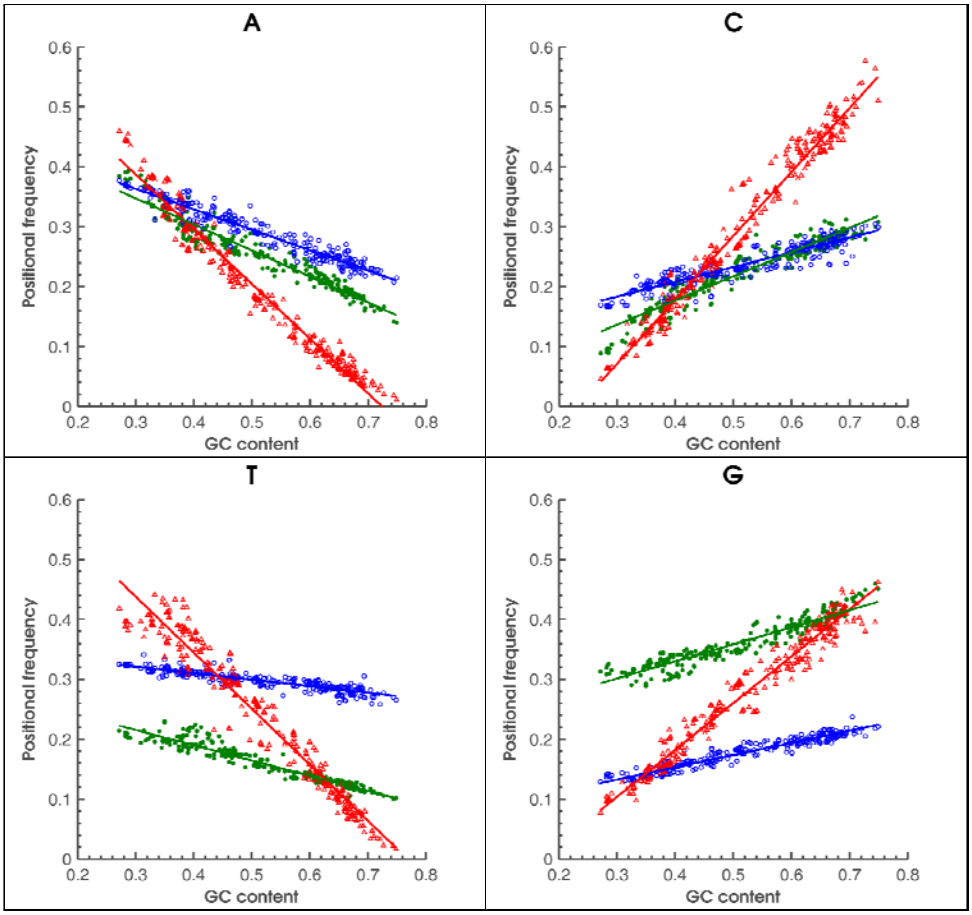


Figure 2.1 Observed frequencies of four nucleotides in the three codon positions (first- green, second- blue, third- red) as functions of genome GC content for 319 bacterial genomes.

Nucleotides G and T have more contrast in frequencies in the first and the second position in comparison with A and C. Frequencies in the third codon position are most sensitive to genome GC content.

Table 2.1 Values of slopes of linear regression lines (such as in Figure 2.1).
Section 1a: slope values for frequencies of nucleotides in the three codon positions for bacterial (B) and archaeal species (A). Section 1b: the same as in section 1a for mesophilic (M) and thermophilic (T) species. Sections 1a and 1b show almost identical sets of slope values for bacterial and mesophilic divisions. Slope values of archaeal and thermophilic divisions are distinctly different.

(a)

Nucleotide type	Archaea/Bacteria	Codon position		
		1	2	3
A	B	-0.43	-0.34	-0.91
	A	-0.50	-0.29	-0.97
C	B	0.40	0.25	1.07
	A	0.38	0.21	1.04
T	B	-0.25	-0.11	-0.93
	A	-0.24	-0.10	-0.86
G	B	0.28	0.20	0.78
	A	0.36	0.19	0.79

(b)

Nucleotide type	Mesophilic/ Thermophilic	Codon position		
		1	2	3
A	M	-0.44	-0.34	-0.92
	T	-0.55	-0.32	-0.92
C	M	0.40	0.25	1.07
	T	0.51	0.25	1.01
T	M	-0.25	-0.11	-0.93
	T	-0.25	-0.15	-0.81
G	M	0.28	0.20	0.78
	T	0.30	0.22	0.72

frequencies of amino acid α observed in corresponding proteomes with respect to the genomic GC contents. To give an example, for alanine with four synonymous codons, predicted frequency f_{GCT} of codon GCT is:

$$f_{GCT} = f_{alanine}(s) \times \left[f'_{GCT} / (f'_{GCT} + f'_{GCG} + f'_{GCC} + f'_{GCA}) \right]$$

Note that the left part of the formula does not change in further iterations (i.e. by substituting thus found f_{GCT} into right part of the equation).

Finally, it was shown that all parameters of the three periodic Markov chain model of a protein coding region could be determined as functions of the set of predicted codon frequencies (Besemer and Borodovsky 1999). A model of non-coding region was defined as the multinomial model, the zero-order Markov model. GC content of non-coding regions is observed to have strong correlation with the genome wide GC content (Figure 2.2). Therefore, nucleotide frequencies observed in a relatively short DNA fragment are accepted as estimates of four parameters of the non-coding region model. Thus parameterized models of protein-coding and non-coding regions are ready for use in a gene finding program such as GeneMark.hmm (Lukashin and Borodovsky 1998; Besemer, Lomsadze et al. 2001).

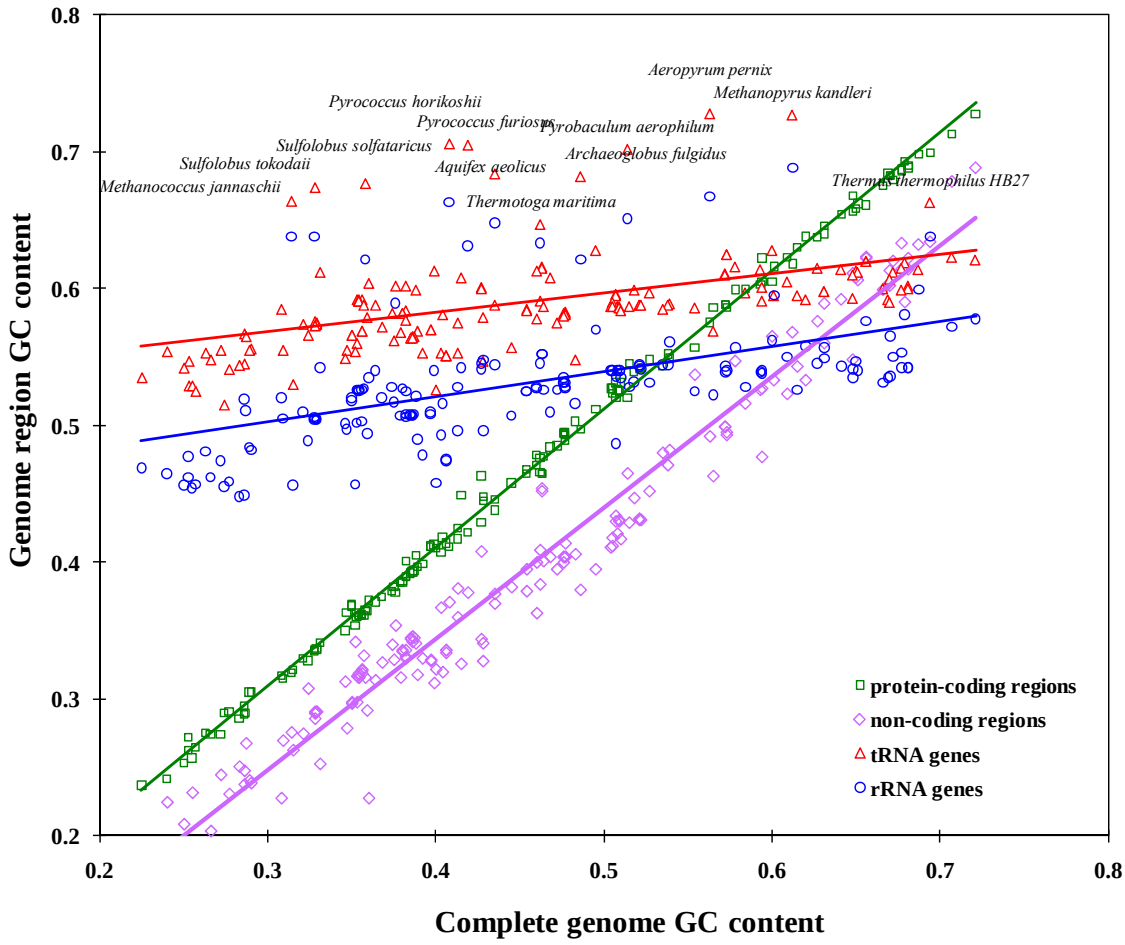


Figure 2.2 Dependence of GC content of genomic functional regions on genome wide GC content.

Protein-coding and non-coding regions were identified in randomly selected 155 bacterial and 16 archaeal genomes by GeneMarkS; tRNA genes by tRNAScan-SE ,while rRNA genes were selected as annotated in RefSeq. Triangles and circles in the top of the figure, with species names, indicate GC content of tRNA and rRNA genes of archaeal thermophiles with higher GC content than tRNA and rRNA genes of mesophilic species.

2.3.2 Refined methods for estimation of parameters of the model of protein coding regions

With hundreds of prokaryotic genomes sequenced and annotated, it is possible to use non-linear (polynomial or logistic) regression to more precisely determine the dependence of codon frequencies on genome GC content. To choose the order of regression polynomial we recall the observed linearity in dependence of frequencies of nucleotides in the three codon positions on genome GC content; product of three linear functions is natural to approximate by the third order polynomial $A + Bs + Cs^2 + Ds^3$; the least squares method is applied to estimate the four coefficients.

A logistic function $f(z) = \frac{1}{1+e^{-z}}, (z = \beta_0 + \beta_1 s)$ could approximate observed codon frequencies scaled with respect to the minimum and maximum values: $f^{scaled} = (f - f^{min})/(f^{max} - f^{min})$, this approach was used earlier (Noguchi, Park et al. 2006). A generalized linear regression function *glmfit* from the MatLab Statistics Toolbox was used to determine β_0 and β_1 parameters from the equation $\ln\left(\frac{f^{scaled}}{1-f^{scaled}}\right) = \beta_0 + \beta_1 s$. For a given *s*, a codon frequency was determined as follows. With $f^{scaled} = \frac{1}{1+e^{-z(s)}}$ predicted codon frequency was determined as $f(s) = f^{scaled} * (f^{max} - f^{min}) + f^{min}$. Frequencies of 64 nucleotide triplets residing in each of other two reading frames could be reconstructed by either one of the two regression approaches outlined above. The three vectors of triplet frequencies thus reconstructed for a short sequence S with respect to its GC content are sufficient for computing parameters of the second order three-periodic Markov chain model, the model of protein-coding region in an unknown genome sequence S came from.

Summarizing the options described above, parameters of the three-periodic second order Markov chain could be determined by several alternative techniques: A/ reconstructing

codon frequencies from predicted nucleotide frequencies in the three codon positions, with subsequent derivation of triplet frequencies in the second and third frame (Besemer and Borodovsky 1999); this technique is called below HAL-99 (the Heuristic ALgorithm); B/ reconstructing codon frequencies by the third order polynomial functions, the rest, for triplet frequencies, is the same as in HAL-99; the C-3 technique; C/ reconstructing frequencies of K-mers, K= 3, 4, 5, 6 in the three frames with the K-order polynomial regression; the K-K techniques; D/ reconstructing frequencies of K-mers, K= 3, 4, 5, 6 in the three frames with the logistic regression; the K-L techniques.

We show examples of typical regression graphs forcodons AAT, GCC, TTG and CGT frequencies observed in bacterial genomes (Figure 2.3); the regression curves were produced by the HAL-99, C-3 and 3-L techniques. Codon AAT is A and T "rich". As a rule, frequencies of 8 out of 64 AT rich codons show monotonous decrease over the whole GC range with a rather small variation in any given GC content (Figure 2.3a). The codon GCC frequency, as well as frequency of other 7 GC rich codons increases as genome GC content grows (Figure 2.3b). Frequencies of codons with mixed composition, such as TTG and CGT (Figure 2.3c,d) show more variation particularly in the mid GC range and task of approximation these frequencies by a function of single variable is more challenging. It was reported that in genomes with the same GC content the differences in codon frequencies correlate with differences in optimal growth temperature, *t* (Lobry and Necsulea 2006). These observations motivate introduction of yet another technique, designated as the C-M technique using approximation of codon frequencies by a function of two variables $f_{XYZ} = A + Bs + Cs^2 + Ds^3 + Et + Fst$, the

sequence GC content, *s*, and the temperature of microbiome habitat, *t*, with parameters determined by multiple regression (Figure 2.4).

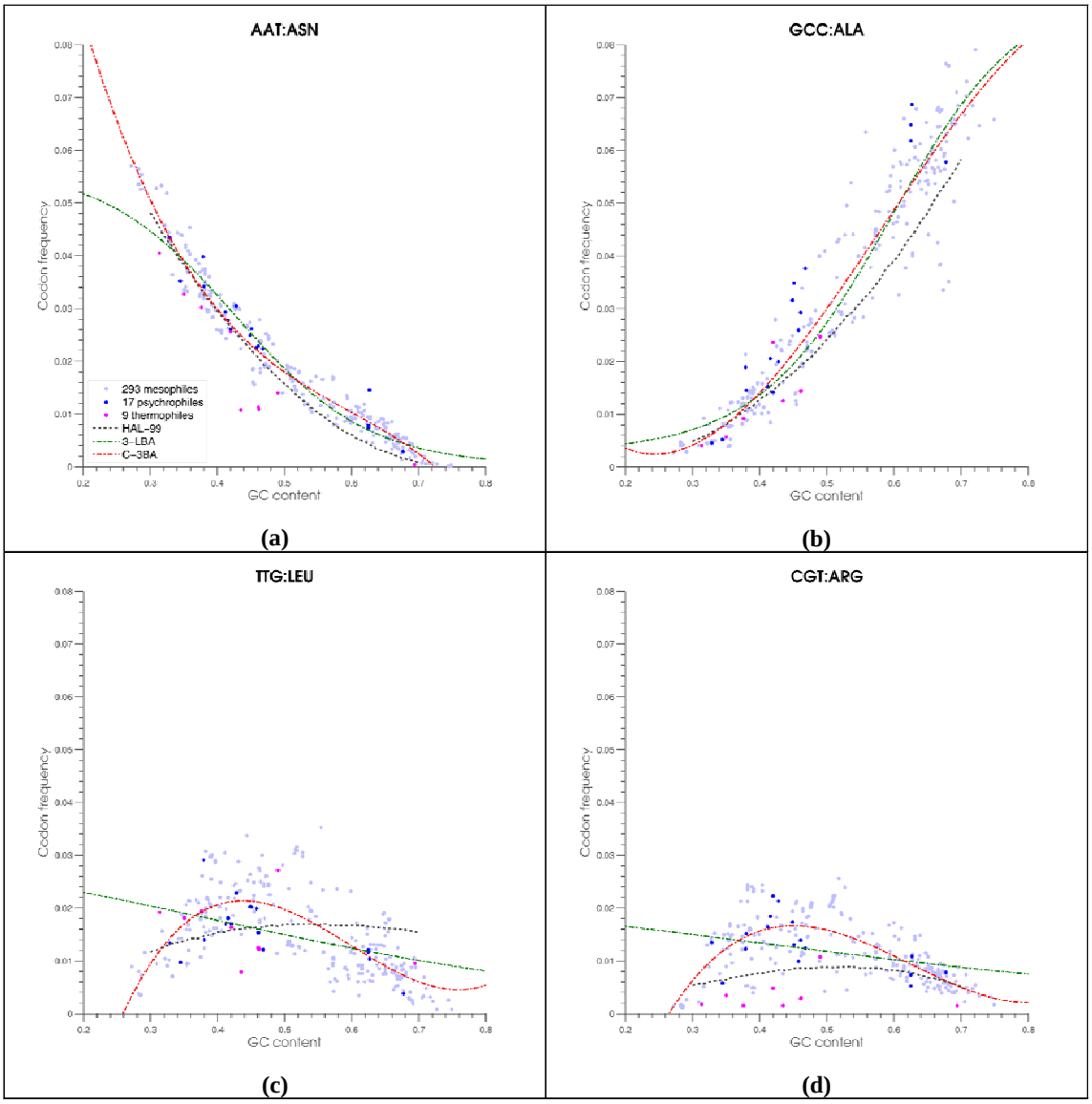


Figure 2.3 Characteristic cases of codon frequency dependence on genome GC content. Each panel shows observed frequencies of a given codon in 319 bacterial genomes. Mesophilic, psychrophilic and thermophilic species are shown as light blue, dark blue and purple dots, respectively. Three techniques of approximating dependence of codon frequency from genome GC content are illustrated: 1999 heuristic model (HAL-99, black dotted line); logistic regression (3-L, green dotted line) and order three polynomial regression (C-3, red dotted line). Plots for 61 codons are available exon.gatech.edu/GeneMark/metagenome/Training/PlotPDF/BAC2D.pdf.

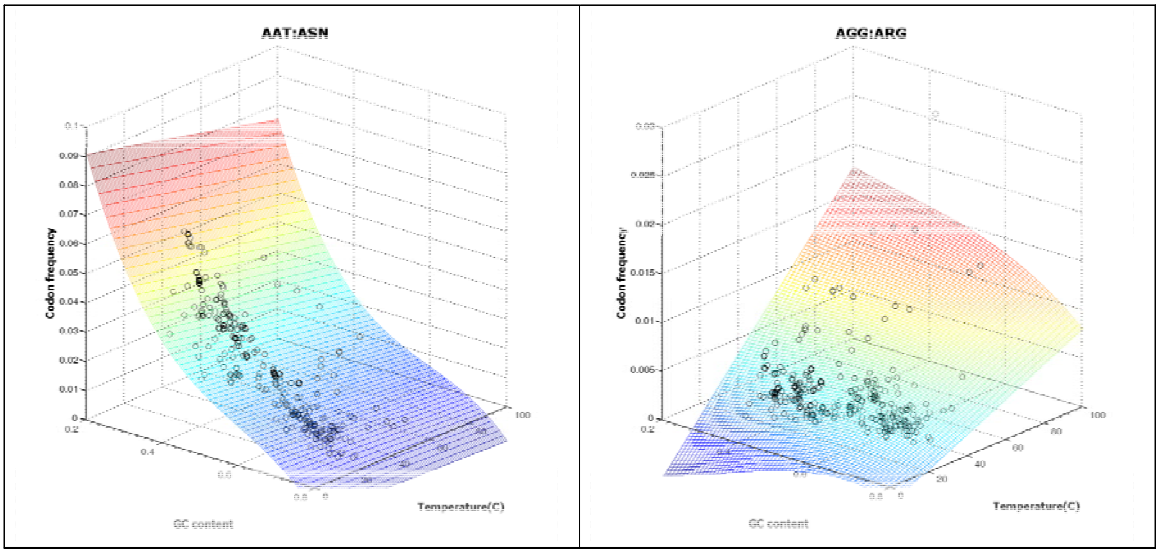


Figure 2.4 Result of multiple regression polynomial fitting of codon frequency as a function of both genomic GC content and optimal growth temperature.
Note that the scales in Z-axes are not the same. Frequency of AAT mostly depends on genomic GC content, adding one more predictor variable explained just additional 1% variance (R^2 value increased from 96% to 97%). Frequency of AGG largely depends on the optimal growth temperature; 30% of variance was explained by the temperature predictor. (R^2 value increased to 31% from 1%). The surface plot indicates a codon frequency change by color, from low (blue) to high (red). Plots for all 61 codons are available at:
exon.gatech.edu/GeneMark/metagenome/Training/PlotPDF/BAC3Dmulti.pdf.

2.3.3 Dual mode of using heuristic models

Linear trends in frequencies of nucleotides in the three codon positions with respect to genome GC content have been observed to be different in bacteria and archaea (Table 2.1a). Therefore, two distinct heuristic models could be built, one for bacterial and another one for archaeal sequences. Still, no pre-processing is needed to identify a domain of life the short sequence fragment represents. The bacterial and archaeal heuristic models can be used in the GeneMark.hmm algorithm simultaneously (Figure 2.5), similarly to the simultaneous use of typical and atypical gene models (Lukashin and Borodovsky 1998). A protein-coding region, if present in the sequence, is supposed to be recognized by either bacterial or archaeal model.

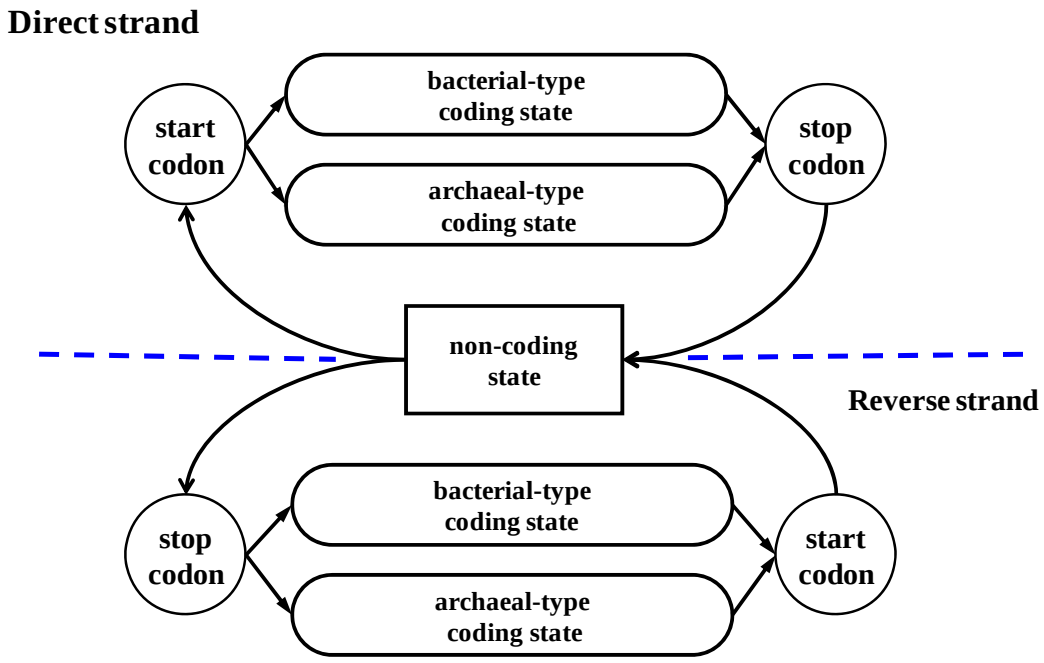


Figure 2.5 Hidden states diagram of the generalized HMM used in the GeneMark.hmm algorithm.
This is the case of using bacterial and archaeal model pair (a similar diagram would be valid for use of mesophilic and thermophilic model pair).

Alternatively, all prokaryotic species could be divided into mesophilic and thermophilic (310 mesophilic and 47 thermophilic in our reference set of sequenced genomes). Then, application of regression analysis of nucleotide frequencies in the three codon positions produced once again two distinct sets of 12 linear functions (Table 2.1b). The two heuristic models (built for mesophiles and thermophiles) could also be used simultaneously in GeneMark.hmm. However, such a dual model seems to be less effective for practical use as the temperature of a microbiome habitat is supposed to be known and one of the models could be chosen *a priori*.

In the Results section we designate the model pairs by suffix BA or TM, e.g. 3-3BA stands for use a pair of bacterial and archaeal models derived by the third order polynomial approximation of triplet frequencies.

2.3.4 Length distributions for partial and complete genes

An average gene length in a prokaryotic genome is about 900nt. In a metagenomic sequence shorter than 900nt, it is more likely to observe a part of a gene than a complete gene. To account for frequent occurrence of partial genes we have to modify a formula for the gene length frequency distribution used in GeneMark.hmm for gene finding in complete genomes. This distribution of whole gene length is approximated by $p(d) = N_c(d/d_c)^2\exp(-d/d_c)$, the γ distribution formula with two parameters (Lukashin and Borodovsky 1998). Also, the length distribution of non-coding regions is approximated by exponential distribution $p(d)=N_n\exp(-d/d_n)$. Parameters, d_c and d_n are estimated by fitting to empirical distributions of gene length in known genomes. It was observed that values of d_c and d_n vary little among different prokaryotic species. Therefore, values of

these parameters in the algorithm were given as default values: d_c = 300 and d_n = 150. The formula for length distribution of protein-coding regions in short metagenomic sequences formula is $p(d) = N_p(d^2+d_cd+2d_c^2)\exp(-d/d_c)$, with parameters N_p and d_c. Corresponding graphs of theoretical and observed length distributions are shown in Figure 2.6. To avoid predicting too short partial genes we have defined an effective 60nt minimum length of a predicted gene by setting $p(l <= 60) = 0$.

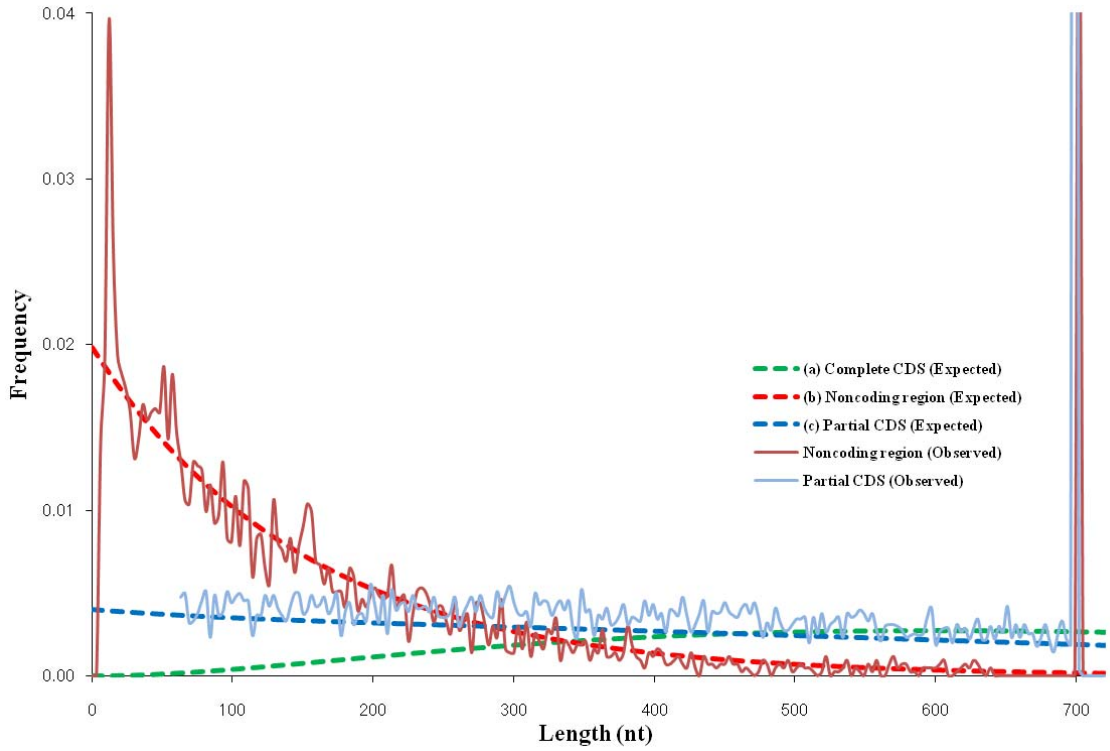


Figure 2.6 Length distributions of coding and non-coding regions observed and expected in 700nt long fragments of E. coli K12 genome.
An average *E. coli* gene length is about 900 nucleotides. Therefore, some of 700nt fragments are 100% coding, hence the peak of frequency of partial CDS length (light blue) at 700nt point. Similarly, the frequency of length of non-coding region has two peaks at 15 and 700nt. (a) Complete CDS length distribution is approximated by function $g(d) = N_c(d/d_c)^2\exp(-d/d_c)$, d_c =300; (b) Noncoding region length distribution is approximated by function $f(d)=N_n\exp(-d/d_n)$, d_n = 150; (c) Partial CDS length distribution is approximated by function $p(d) = N_p(d^2+d_cd+2d_c^2)\exp(-d/d_c)$, d_c =300.

2.4 Results

2.4.1 Choice of parameters of length distributions

To analyze how accuracy of GeneMark.hmm depends on d_c and d_n values we used sets of 700nt long fragments of *E. coli* and *B. subtilis* genomes; the model used in the runs was the C-3BA one. Sensitivity (*Sn*) and Specificity (*Sp*) were determined by comparison of gene predictions with fragments annotation. A prediction was accounted as a true positive if locations of the predicted and annotated 3' ends matched within the sequence or for partial genes without 3'ends there was a match between predicted and annotated reading frames. The values of d_c could vary from 100 to 800, while values of d_n varied from 100 to 300. Particularly, dependence of *Sn* and *Sp* for d_c = 800 while d_n is varying between 100 to 300 is shown in Figure 2.7 by blue line; similarly, dependence of *Sn* and *Sp* for d_n = 100 while d_c is changing from 100 to 800 is shown by purple line. The d_c, d_n setting used for analysis of complete genomes (300,150) is indicated by red dot. Combining larger d_c (800) and smaller d_n (100) leads to a substantial increase of *Sp* and a slight decrease of *Sn*. This result is due to decrease in number of predicted short genes, many of them not matching annotation. To facilitate comparison of average values $S = (Sn + Sp)/2$, produced by the program runs with different d_c, d_n values the S constant level lines (with a slope of -1) were plotted in Figure 2.7a,b. Performance (*Sn, Sp*) of MetaGene and MetaGeneAnnotator (with default parameters) was depicted for each of two genomes as well; one can see that the performance is high, though it can be outperformed, especially in the *E. coli*, by GeneMark.hmm with a wide range of parameters d_c, d_n. As the result of the modeling we have used d_c=800, d_n = 100 in further analysis of artificial and real metagenomic sequences.

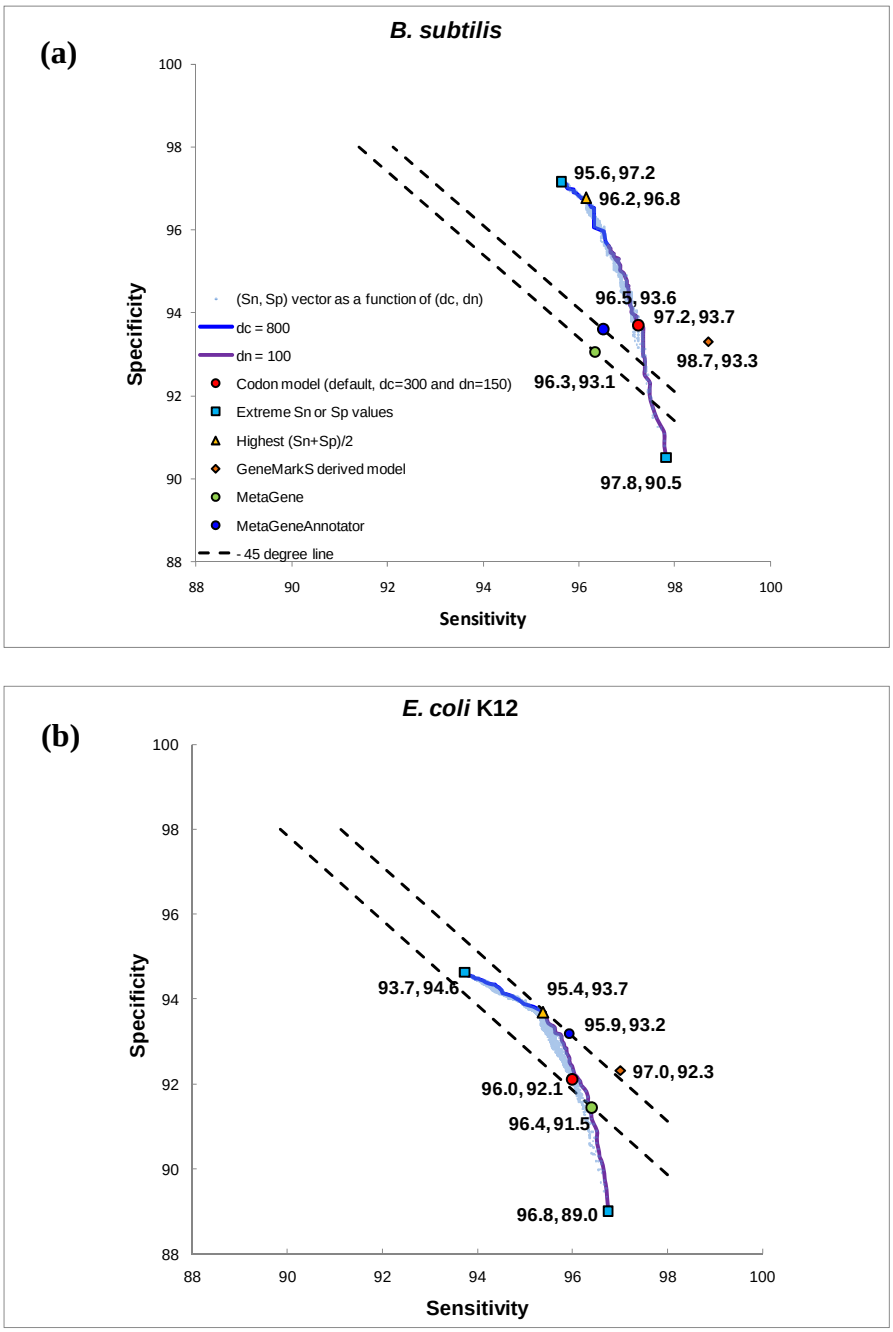


Figure 2.7 Values of Sn and Sp obtained upon variations of parameters dn and dc.
Light blue dots represent Sn and Sp values obtained for each of 1491 combinations of (d_n, d_c) parameters. Blue and purple lines correspond to variation of d_n with d_c=800 and variation of d_c with d_n=100, respectively. Red dots correspond to (d_n, d_c) setting (150, 300) which is used by default for complete genomes. Also shown are the highest Sn and the highest Sn (blue squares), the highest (Sn+Sp)/2 (yellow triangles). Use of pair of models, the native model (derived by the GeneMarkS from a complete genome) and the heuristic model HAL-99, produced the Sn and Sp values shown by orange diamonds. The Sn and Sp of the MetaGene and MetaGeneAnnotator predictions are shown by green and blue dots, respectively.

2.4.2 Tests on sequences with fixed length

We used the GeneMark.hmm program with the pairs of heuristic models, bacterial and archaeal (or mesophilic and thermophilic) derived by methods described above to analyze sequence fragments with fixed length, from 50 genomic sequences (the list is given in Supplementary Table 2). All models were tested on sets of fragments with length of 400nt and 700nt, however, the models with highest performance were tested on sets of fragments with shorter (down to 63nt) and longer (up to 1100nt) length. Performance characteristics of different models are shown in Table 2.2 (with more details provided in Supplementary Table 3, Supplementary Table 4, Supplementary Table 5 and Supplementary Table 6). Observed values of $(Sn + Sp)/2$ clustered between 94.5% and 96.5% for 700nt long fragments and between 93.5% and 96.0% for 400nt long fragments. Interestingly, among the triplet based models, C-3BA, C-3MT, 3-3BA, 3-LBA, the codon frequency derived models, C-3BA and C-3MT, demonstrated higher performance than 3-3BA and 3-LBA models where frequencies of triplets as functions of GC content are independently approximated in each frame. Use of higher order Markov models: the third order, 4-4BA, the fourth order, 5-5BA, and the fifth order, 6-6BA and 6-LBA, resulted in similar performance, with differences in $(Sn + Sp)/2$ values smaller than 0.3%; this performance level is comparable to performance of the second order models C-3BA and C-3MT. Still, a slightly higher $(Sn + Sp)/2$ for 700nt and 400nt long fragments was achieved with the use of 6-LBA heuristic model containing a pair of the fifth order model, bacterial and archaeal, with parameters obtained by logistic regression approximation of hexamer frequencies. Note that the MetaGene authors found performance of MetaGene on 700 nt fragments comparable to performance of

GeneMark.hmm with HAL-99 model (Noguchi, Park et al. 2006; Suppl. Table 3). This result corresponds to our observations as well (Table 2.2).

Table 2.2 Accuracy of gene prediction in 700nt and 400nt long fragments from 50 genomic sequences (listed in Suppl. Table 2).
Values of length distribution parameters: d_n=100 and d_c = 800.

(a)

		700nt		
Program	Model	Sensitivity	Specificity	(Sn+Sp)/2
GeneMark.hmm	HAL-99	94.93	94.28	94.61
	C-3BA	96.84	95.17	96.01
	C-3MT	96.86	95.04	95.95
	C-MBA	97.00	93.77	95.39
	3-3BA	96.51	94.18	95.35
	3-LBA	96.69	94.19	95.44
	4-4BA	97.23	94.83	96.03
	5-5BA	97.25	94.91	96.08
	6-6BA	97.04	94.99	96.02
	6-LBA	97.42	94.89	96.16
MetaGene		97.57	92.36	94.97
MetaGeneAnnotator		97.49	93.60	95.55

(b)

		400nt		
Program	Model	Sensitivity	Specificity	(Sn+Sp)/2
GeneMark.hmm	HAL-99	93.81	93.38	93.59
	C-3BA	96.24	94.80	95.52
	C-3MT	96.32	94.72	95.52
	C-MBA	96.34	93.31	94.83
	3-3BA	95.64	93.85	94.74
	3-LBA	95.97	93.77	94.87
	4-4BA	96.70	94.57	95.63
	5-5BA	96.75	94.66	95.70
	6-6BA	96.49	94.77	95.63
	6-LBA	96.99	94.63	95.81
MetaGene		97.22	91.08	94.15
MetaGeneAnnotator		97.15	92.35	94.75

The use of models utilizing higher order oligonuclotides brought in a marginal improvement of $(Sn + Sp)/2$ for gene prediction in 400nt and 700nt fragments in comparison with the codon based models, e.g. C-3BA and C-3MT (Table 2.2, Supplementary Table 3, Supplementary Table 4, Supplementary Table 5 and Supplementary Table 6). This observation corroborates observations of other authors that use of the fifth order Markov chains and/or di-codon frequencies led to a slight increase in gene prediction accuracy (Noguchi, Park et al. 2006; Hoff, Tech et al. 2008; Noguchi, Taniguchi et al. 2008). In order to determine accuracy of gene prediction in fragments with length other than 400nt and 700nt, the particular lengths that have been used in tests by several authors, we have extended the range of test sets derived from the 50 genomes to 11 other fragment lengths including ones shorter than 400nt (Table 2.4). Here, in comparison of MetaGene, MetaGeneAnnotator with GeneMark.hmm using HAL-1999, C-3BA and 6-LBA models we see that the 6-LBA models perform marginally better in terms of *Sn* and *Sp* average. Still MetaGene shows higher *Sn* for all the 13 test sets, while C-3BA shows higher *Sp* for fragment length longer than 200nt. For better visualization we show the programs performance as functions of fragment length for the sets with fragment length ≥100nt (Figure 2.8 and Table 2.3). Notably, since the second order C-3BA model is very close to the 6-LBA model in terms of performance, we use the C-3BA model in several applications discussed below along with the 6-LBA model (Table 2.2, Table 2.4, Supplementary Table 3, Supplementary Table 4, Supplementary Table 5 and Supplementary Table 6).

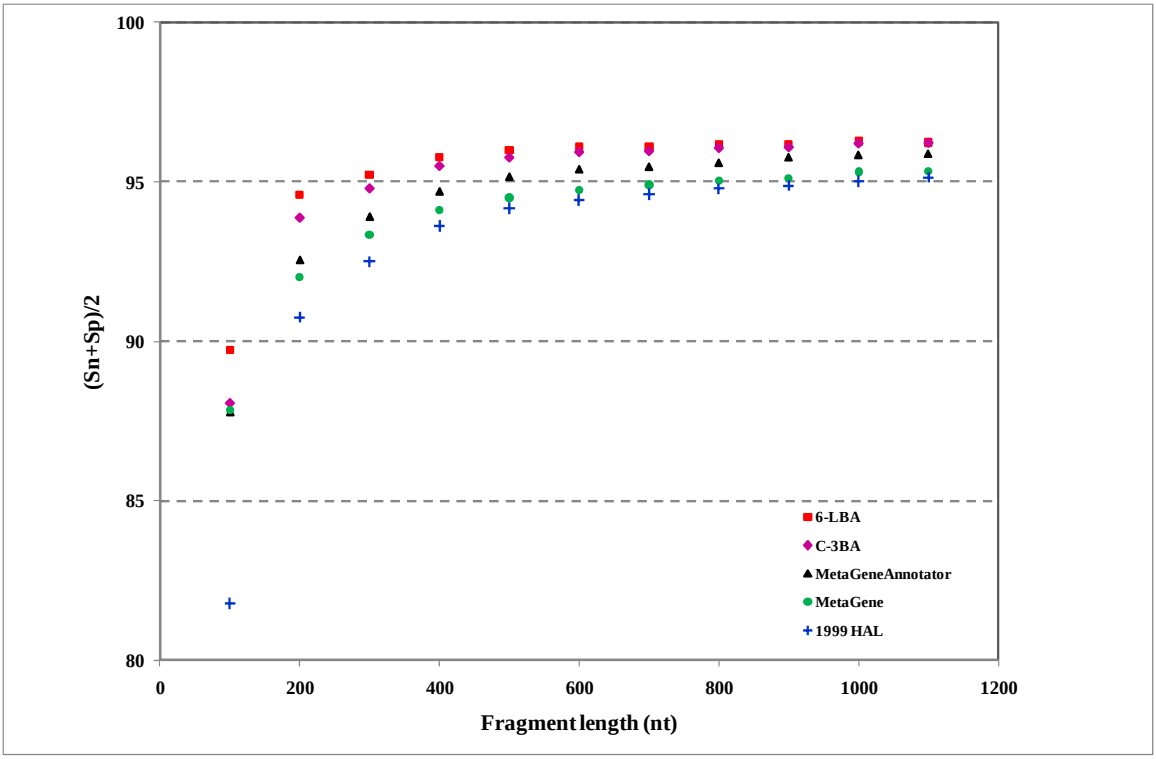


Figure 2.8 Gene prediction accuracy of GeneMark.hmm with three different heuristic models as well as MetaGene and MetaGeneAnnotator observed on the sets of sequence fragments from 50 genomes with length from 100nt to 1100nt.

Table 2.3 Standard deviation of five different methods in Figure 2.8

Length	1999 HAL	MetaGene	MetaGene Annotator	C-3BA	6-LBA
100	6.8	3.8	3.8	4.4	3.8
200	4.4	2.7	2.6	2.5	2.1
300	3.5	2.2	2.2	2.0	1.9
400	2.9	2.0	2.0	1.8	1.7
500	2.6	1.9	1.8	1.7	1.7
600	2.5	1.9	1.8	1.7	1.7
700	2.2	1.7	1.8	1.6	1.6
800	2.1	1.8	1.7	1.5	1.6
900	2.1	1.7	1.7	1.6	1.6
1000	2.0	1.6	1.6	1.5	1.5
1100	1.9	1.7	1.7	1.6	1.6

Table 2.4 Gene prediction accuracy of GeneMark.hmm with three different heuristic models as well as MetaGene and MetaGeneAnnotator observed on the sets of sequence fragments from 50 genomes with length from 72nt to 1100nt.
The best numbers are bold.

Length		1999 HAL		MetaGene		MetaGeneAnnotator		C-3BA		6-LBA	
72	**Sn**	64.5	72.8	n/a	n/a	**84.2**	83.1	77.8	81.7	81.2	**84.0**
	Sp	81.1		n/a		82.1		85.5		**86.8**	
96	**Sn**	77.0	80.8	n/a	n/a	**90.6**	87.3	85.9	87.3	88.6	**89.1**
	Sp	84.6		n/a		84.0		88.7		**89.6**	
100	**Sn**	78.4	81.8	**91.2**	87.8	90.9	87.8	87.0	88.1	89.4	**89.7**
	Sp	85.1		84.5		84.6		89.2		**90.0**	
200	**Sn**	90.7	90.8	**95.7**	92.0	95.6	92.5	94.3	93.9	95.6	**94.6**
	Sp	90.9		88.3		89.5		93.4		**93.6**	
300	**Sn**	92.7	92.5	**96.8**	93.3	96.7	93.9	95.5	94.8	96.4	**95.2**
	Sp	92.3		89.9		91.1		**94.1**		94.0	
400	**Sn**	93.9	93.6	**97.3**	94.1	97.2	94.7	96.3	95.5	97.0	**95.8**
	Sp	93.3		90.9		92.2		**94.7**		94.5	
500	**Sn**	94.4	94.2	**97.5**	94.5	97.4	95.2	96.6	95.8	97.2	**96.0**
	Sp	93.9		91.5		92.9		**95.0**		94.8	
600	**Sn**	94.8	94.4	**97.6**	94.7	97.5	95.4	96.9	95.9	97.5	**96.1**
	Sp	94.0		91.9		93.3		**95.0**		94.7	
700	**Sn**	95.0	94.6	**97.6**	94.9	97.5	95.5	96.9	96.0	97.4	**96.1**
	Sp	94.2		92.2		93.4		**95.0**		94.8	
800	**Sn**	95.2	94.8	**97.7**	95.0	97.6	95.6	97.0	96.1	97.6	**96.2**
	Sp	94.3		92.4		93.6		**95.1**		94.8	
900	**Sn**	95.4	94.9	**97.7**	95.1	97.7	95.8	97.1	96.1	97.6	**96.2**
	Sp	94.4		92.5		93.8		**95.1**		94.7	
1000	**Sn**	95.5	95.0	**97.9**	95.3	97.8	95.8	97.2	96.2	97.7	**96.3**
	Sp	94.5		92.8		93.9		**95.2**		94.8	
1100	**Sn**	95.7	95.1	**97.8**	95.3	97.7	95.9	97.3	96.2	97.7	**96.2**
	Sp	94.5		92.9		94.0		**95.2**		94.7	

2.4.3 Inferring origin of genes and sequence fragments

Results of gene prediction in of the 50 complete prokaryotic genomes (Supplementary Table 7, Supplementary Table 8) demonstrated clearly that, as a rule, a vast majority of genes in a given bacterial (archaeal) genome was predicted by the bacterial (archaeal) model. Similarly, a vast majority of genes in a thermophilic (mesophilic) genome were predicted by thermophilic (mesophilic) model. Interestingly, for the thermophilic bacteria *Thermotoga maritima* (with optimal growth temperature 80C) the archaeal model predicted 3137 out of a total of 3225 fragmented genes, corroborating the findings made in the original *T. maritima* genome paper (Nelson, Clayton et al. 1999) of massive horizontal influx of genes transferred from archaeal species (Zavala, Naya et al. 2002). On the other hand, a vast majority of genes in *Methanosarcina acetivorans,* identified in many sources as mesophilic archaea, were predicted by the thermophilic model. This result corresponds to observations that *M. acetivorans* is able to live in deep sea hydrothermal vents. Similar observations were made for bacteria *Aquifex aeolicus* (Basak, Banerjee et al. 2004) living in high temperature as well as for low temperature archaeal species *Haloarcula*, *Halobacterium* and *Methanosphaera*.

Similarly, in the case of metagenomic sequences, a run of GeneMark.hmm with bacterial and archaeal model pair produced not only predicted genes but also an indication of a likely origin of each gene. In short fragments one rarely seen more than one gene per fragment, therefore, a gene characterization is extended to the whole fragment. Rare cases when there are several genes in a metagenomic fragment each predicted by different model is worthwhile to set aside as candidate cases for study of horizontal gene transfer. All around, in the test set of 700nt long fragments, with a total of 31,584 archaeal (136,210 bacterial) fragments, GeneMark.hmm with C-3BA model misclassified

2,757 fragments as bacterial-type (16,284 fragments as archaeal-type), thus archaeal fragments were identified correctly in 91.27% of cases and bacterial fragments were identified correctly in 88.04% of cases (Supplementary Table 7, column C-3BA). Similar analysis for set of 400nt long fragments resulted in 89.92% coorect predictions for archaea and 87.26% for bacteria (Supplementary Table 8, column C-3BA). Note that a domain classification within a metagenomic gene finder was first proposed by Noguchi et al. The difference with their method is rather technical as the domain recognition in GeneMark.hmm is embedded in the run of Viterbi algorithm as an assignment of the most likely type of a hidden state for predicted coding region, bacterial or archaeal.

2.4.4 Analysis of sequences from human and mouse gut microbiomes

We used GeneMark.hmm with C-3BA model to predict genes in metagenomic sequences from two human and five mouse gut microbiomes (Table 2.5). In these sequence sets we have identified 11,865 genes not annotated earlier. Protein products of 1,984 genes (in human samples) and 3,435 genes (in mouse samples) had similarity to known proteins detectable by BLASTP with E-value threshold 10^{-5}. Protein functions that could be assigned to the 50 longest genes predicted in the gut microbiomes derived sequences are listed in Supplementary Table 9. A relative proportion of new genes in the mouse gut metagenomic sequences is about three times higher than in human ones; the mere numbers are about or more than 50% of the number of initially annotated genes. Interestingly, 17% (15%) of the metagenomic sequences in human subject 7 (8) could be mapped to known genomes of bacteria and archaea (Table 2.6, Table 2.7 Table 2.8) by the BLASTN search with E-value threshold 10^{-13}. Interestingly, in the metagenomic

sequences from mice guts we were not able to identify DNA sequence fragments highly similar to a sequence in already sequenced genomes (with threshold 10^{-13}).

Table 2.5 Results of analysis of metagenomic sequences from human and mouse gut microbiomes.
Annotation coordinates were retrieved from JGI IMG/M database (Markowitz, Ivanova et al. 2008). Note, that the total numbers of genes annotated in JGI IMG/M are different than the numbers of genes given in original publications (Gill, Pop et al. 2006). This is because JGI IMG/M used YACOP, a combination of several gene finding methods, namely Critica, Glimmer and ZCURVE (Tech and Merkl 2003), while BLASTX and BLASTP were used in original publications to identify genes in metagenomic sequences of human and mouse microbiomes. Annotation was not readily available in the original publications. * Percentage values are computed with respect to the numbers of annotated genes.

Methods	Microbiome size (bp)	# annotated	# Predicted	# Missed	% Missed *	# Novel	% Novel *	% (Missed + Novel)/2	% of novel that have hit to nr
human_sub7									
MetaGene	15,817,685	20523	22271	893	4.4	2641	11.9	8.1	34.6
MetaGeneAnnotator			22164	755	3.7	2396	10.8	7.2	40.5
GM.hmm with C-3BA model			21941	730	3.6	2148	9.8	6.7	40.7
human_sub8									
MetaGene	20,486,813	25980	27750	1223	4.7	2993	10.8	7.7	38.2
MetaGeneAnnotator			27707	971	3.7	2698	9.7	6.7	41.7
GM.hmm with C-3BA model			27589	840	3.2	2449	8.9	6.1	45.3
mouse_lean1									
MetaGene	2,234,664	2935	4579	244	8.3	1888	41.2	24.8	40.6
MetaGeneAnnotator			4417	216	7.4	1698	38.4	22.9	44.0
GM.hmm with C-3BA model			4279	236	8.0	1580	36.9	22.5	47.6
mouse_lean2									
MetaGene	2,133,081	2782	4279	296	10.6	1793	41.9	26.3	32.1
MetaGeneAnnotator			4152	265	9.5	1635	39.4	24.5	35.7
GM.hmm with C-3BA model			3950	264	9.5	1432	36.3	22.9	43.9
mouse_lean3									
MetaGene	2,143,888	2793	4262	202	7.2	1671	39.2	23.2	38.7
MetaGeneAnnotator			4198	188	6.7	1593	37.9	22.3	42.8
GM.hmm with C-3BA model			3971	195	7.0	1373	34.6	20.8	47.0
mouse_ob1									
MetaGene	2,359,017	3051	4698	218	7.1	1865	39.7	23.4	38.8
MetaGeneAnnotator			4626	196	6.4	1771	38.3	22.4	43.2
GM.hmm with C-3BA model			4432	213	7.0	1594	36.0	21.5	47.7
mouse_ob2									
MetaGene	1,841,347	2331	3675	192	8.2	1536	41.8	25.0	37.2
MetaGeneAnnotator			3599	172	7.4	1440	40.0	23.7	42.8
GM.hmm with C-3BA model			3444	176	7.6	1289	37.4	22.5	50.4

Table 2.6 Summary of the BLASTn results for DNA sequence queries from metagenomic sequences to nr database (with E-value better than 1e-13).

Data source: gut community	# of DNA fragments	# of those with hit to nr (E value <10^-13)
Human subject 7	10411	1768
Human subject 8	12020	1790
Mouse lean 1	2781	0
Mouse lean 2	2615	0
Mouse lean 3	2678	0
Mouse obesity 1	2946	0
Mouse obesity 2	2219	0

Table 2.7 Top 10 most frequent microbes with complete genomes sequenced matching queries from metagenomic sample of gut microbiome of Human subject 7 (with E-value better than 1e-13).

# of sequences	Species	Temperature range	Optimal temperature
458	*Bifidobacterium adolescentis ATCC 15703 DNA, complete genome*	Mesophilic	37C
381	*Methanobrevibacter smithii ATCC 35061, complete genome*	Mesophilic	37-40C
252	*Bifidobacterium longum DJO10A, complete genome*	Mesophilic	37-41C
195	*Eubacterium rectale ATCC 33656, complete genome*	Mesophilic	n/a
128	*Bifidobacterium longum subsp. infantis ATCC 15697, complete genome*	Mesophilic	37-41C
47	*Bifidobacterium longum NCC2705, complete genome*	Mesophilic	37-41C
21	*Clostridium difficile R20291 complete genome*	Mesophilic	n/a
7	*Eubacterium eligens ATCC 27750, complete genome*	Mesophilic	n/a
5	*Eubacterium eligens ATCC 27750 plasmid, complete sequence*	Mesophilic	n/a
4	*Lactococcus lactis subsp. lactis Il1403, complete genome*	Mesophilic	40C

Table 2.8 Top 10 most frequent microbes with complete genomes sequenced matching queries from metagenomic sample of gut microbiome of Human subject 8 (with E-value better than 1e-13).

# of sequences	Species	Temperature range	Optimal temperature
503	*Eubacterium rectale ATCC 33656, complete genome*	Mesophilic	n/a
413	*Methanobrevibacter smithii ATCC 35061, complete genome*	Mesophilic	37-40C
202	*Bifidobacterium longum DJO10A, complete genome*	Mesophilic	37-41C
93	*Bifidobacterium longum subsp. infantis ATCC 15697, complete genome*	Mesophilic	37-41C
30	*Clostridium difficile R20291 complete genome*	Mesophilic	n/a
30	*Bifidobacterium longum NCC2705, complete genome*	Mesophilic	37-41C
22	*Clostridium difficile 630 complete genome*	Mesophilic	37C
19	*Eggerthella lenta DSM 2243, complete genome*	Mesophilic	n/a
17	*Lactococcus lactis subsp. lactis Il1403, complete genome*	Mesophilic	40C
16	*Streptococcus pyogenes MGAS2096, complete genome*	Mesophilic	n/a

However, for less stringent threshold 10^{-5} we observed in each mouse gut metagenomic sample dozens of fragments with similarity to genomes of known species. Typical situations that are prone to errors in annotation are illustrated in Figure 2.9: short genes could be missed (Figure 2.9a). Some genes could be omitted due to artifacts, such as erroneous extension of the 5' end of a gene to the longest possible start (Figure 2.9b); such an extension may overlap a real gene in the opposite strand and this real gene will be missed in annotation.

The whole set of gene prediction is available at:

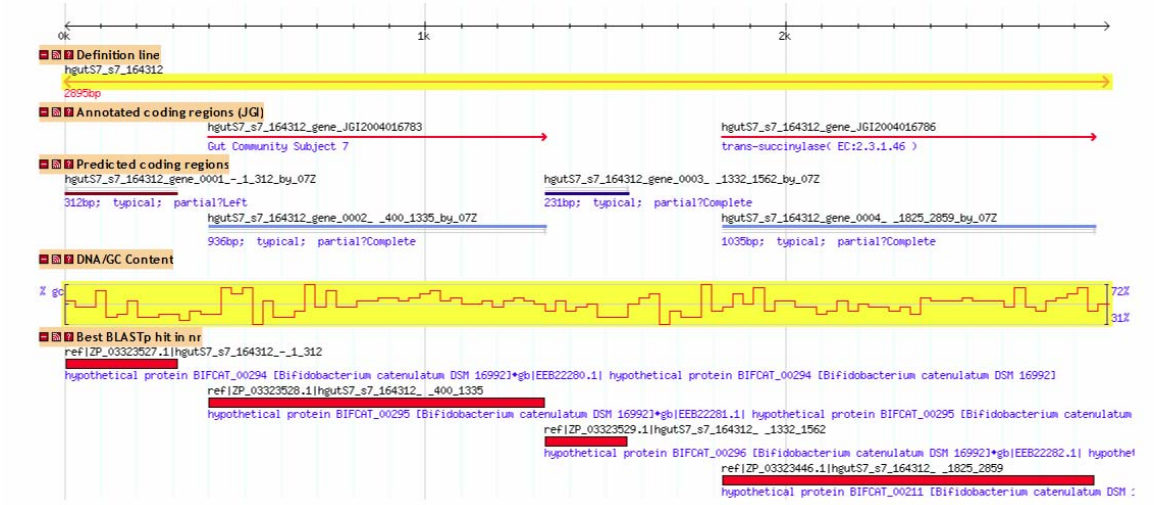
http://exon.gatech.edu/GeneMark/metagenome/database.

It was also visualized in a genome browser utilizing the GBrowse program (Stein, Mungall et al. 2002).

2.4.5 Web interface

We have designed a web site providing an access to the new program for gene prediction in metagenomes: http://exon.gatech.edu/GeneMark/metagenome. For reference purposes we have also provided an interface to the database of genome wide codon frequencies observed in genomes used in the training set.

(a)



(b)

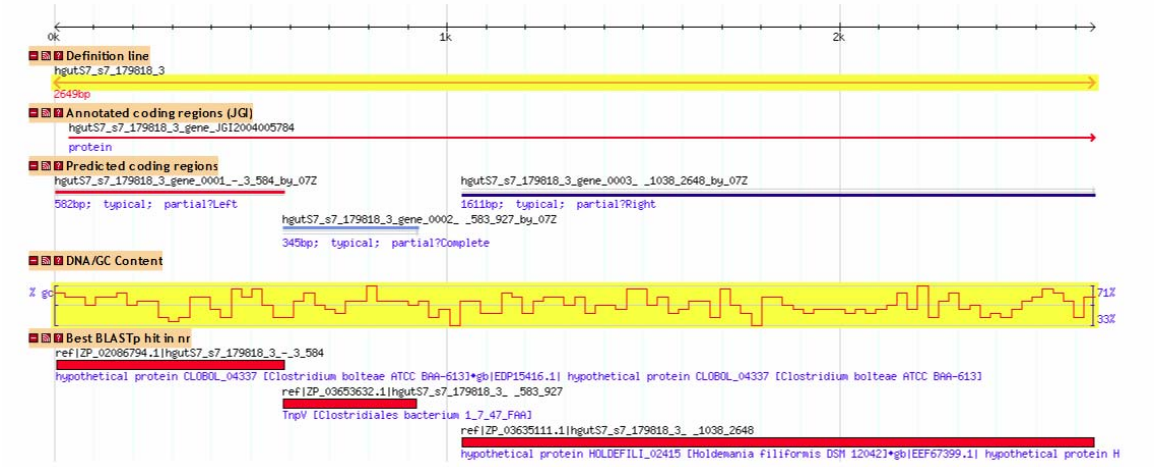


Figure 2.9 Genome Browser view for two sequences from subject 7 human microbiome. The C-3BA model was used to predict coding regions.
(a) The first and third genes shown in panel "Predicted coding regions" were not previously annotated. Protein products of both predicted genes have sequence similarity to proteins in the nr database with E-value of 8e-44 and 2e-35, respectively. (b) In 2,649nt microbiome sequence, a single partial gene was annotated in positive strand in frame +3, starting from nucleotide position 39. New to annotation, three genes were predicted in frames -3, +1 and +3 respectively. Sequences analyzed can be found in Microbiome DB:
exon.gatech.edu/cgi-bin/gbrowse/microbiome_human_sub7/?name=hgutS7_s7_164312;
exon.gatech.edu/cgi-bin/gbrowse/microbiome_human_sub7/?name=hgutS7_s7_179818_3.

2.5 Discussion

Back in 1999 upon analysis of 17 prokaryotic genomes we have determined that genome wide 61 codon frequencies could be approximated by functions of genome wide nucleotide frequencies (Besemer and Borodovsky 1999) the functions of a single parameter, genomic GC content. This critical observation strongly suggested that genomic GC content is the major factor influencing genome wide codon frequencies, the codon usage pattern. It was a hypothesis formulated upon introduction of the heuristic models (Besemer and Borodovsky 1999) and received further support by results independently obtained by other authors (Knight, Freeland et al. 2001; Chen, Lee et al. 2004). The major focus of the current study is on further developing the heuristic models and on their applications to gene finding in metagenomic sequences. Therefore, we had to leave aside intriguing questions on i/ possible evolutionary mechanisms that formed the dependence of codon usage pattern on genome GC content and ii/ how could this dependence evolve differently in the domains of bacteria and archaea or in the classes of mesophilic and thermophilic species.

Notably, both divides, either by phylogeny (bacteria vs archaea) or by the optimal growth temperature (mesophiles vs thermophiles), have produced similar results in terms of accuracy of gene finding in short sequences. Use of the bacteria and archaeal model pair is a natural choice since the origin of a short sequence is not known *a priori*. The second pair of models, mesophilic and thermophilic, may have less frequent use since the temperature of microbiome habitat is known and the model can be chosen *a priori*.

The ability to identify a sequence origin in terms of bacterial or archaeal domain appears to be an added value benefit since on the algorithm automatically identifies the model,

bacterial or archaeal, which fits best the gene sequence and "is attached to" the most likely type of a hidden state. Domain classification was shown to be correct for 88.04% of 700nt long bacterial fragments and for 91.27% of 700nt long archaeal fragments (Supplementary Table 7, Supplementary Table 8). Notably, genes horizontally transferred between the two domains should be responsible for a fraction of misclassification errors.

The results indicate that gene prediction in fragmented sequences of prokaryotic genomes has the same rate of success as in complete prokaryotic genomes. This result is rather surprising as the complete genomes provide a context for each individual sequence fragment and offer much larger sets of sequence data for model training. However, short sequences as targets for gene prediction have some advantages as well. Most of prokaryotic genomes are heterogeneous in terms of GC content. Still, parameters of a conventional model used in a genome wide gene finder are defined for the genome as a whole and the accuracy may slightly suffer in regions whose local GC content deviates from the average one. Derivation of the model parameters for each short sequence individually, is it is done or metagenomic sequences, is likely to tune up parameters for the local GC content and, thus, partially compensate for the insufficient training data.

Existence of a difference between GC content of protein-coding and non-coding regions is a well known fact. However, the nearly constant value of this difference among genomes ranging wide in GC content is an interesting observation (Figure 2.2). Notably, RNA genes have been observed to be uniformly GC rich regardless of genome GC content (Figure 2.2); hence, tRNA genes could be easily detected in AT rich genomes as local regions with a sharp GC content elevation. GC content of protein-coding genes does not correlate with temperature of the species habitat. Still, it is the RNA genes that show

temperature dependent composition. RNA genes in genomes of thermophilic species (genomes that could be either AT or GC rich) have a significantly higher GC content than RNA genes in genomes of mesophilic species (Figure 2.2). A temperature effect on composition of protein coding genes is more subtle and reveals itself in comparison of trends of changes of frequencies of nucleotides in the three codon positions in mesophilic and thermophilic species. Similarly to inferring a domain of origin, bacterial or archaeal, for a gene within the gene finding algorithm with bacterial and archaeal model pair, a pair of heuristic models derived for mesophilic and thermophilic species could be used to for inferring mesophilic or thermophilic origin for an individual gene.

We should mention that the sets of bacterial and mesophilic species used in this study well overlap each other; 301 out of 319 species in the bacterial set are mesophilic. Hence, bacterial and mesophilic protein-coding regions exhibit a similar dependence of frequencies of nucleotides in the three codon positions on genome GC content (Table 2.1). On the other hand, although the set of 38 archaeal species contains 23 thermophiles and overlaps significantly with the set of 47 thermophilic species in this study, most archaeal and thermophilic regression slope coefficients (Table 2.1) are distinctly different.

We should note that frameshifts in protein coding regions, caused by sequencing errors, are more frequent in metagenomes than in complete genomes. It was shown (Hoff 2009) that performance of all current methods for metagenome gene finding including GeneMark.hmm with the original HAL-99 models is sensitive to presence of frameshifts. The new heuristic models make no exception and sensitivity to sequence errors has roughly the same pattern as one already reported for HAL-99 (Hoff 2009). Additionally,

as a separate project we have developed a new algorithm and software tool for frameshift identification (Antonov and Borodovsky 2010) that could be combined with the heuristic models and used for frameshift detection in metagenomic sequences.

In conclusion, we should say that we have presented here methods of reconstruction of codon and oligomer frequencies that have led to new heuristic models for gene finding in short sequences. We have shown that use of the new models in GeneMark.hmm resulted in more accurate gene predictions than use of developed earlier heuristic models HAL-99. The gene prediction accuracy was shown to be higher than that of MetaGene and MetaGeneAnnotator (Table 2.2).

The HAL-99 models have been used in gene prediction and annotation since 1999. They were used in *ab initio* prokaryotic and eukaryotic gene finders GeneMarkS and GeneMark-ES to initiate unsupervised training for complete and nearly complete genomes (Besemer, Lomsadze et al. 2001; Lomsadze, Ter-Hovhannisyan et al. 2005; Ter-Hovhannisyan, Lomsadze et al. 2008). Particularly, HAL-99 were used in *ab initio* gene prediction and annotation in viral genomes (Mills, Rozanov et al. 2003) and in metagenomic sequences at the pipeline of DOE Joint Genome Institute.

* This chapter was part of the following publication (Zhu, Lomsadze et al. 2010):

Zhu W., Lomsadze A. and Borodovsky M. (2010).

ab initio Gene Identification in Metagenomic Sequences.

Accpeted, Nucleic Acids Research

CHAPTER 3 GeneMarkS Plus for Gene Prediction in Complete Prokaryotic Genomes

Abstract

While the accuracy of gene prediction programs has reached 95% or above in prokaryotic genomes, several aspects of this problem have to be addressed to find the integrated solution and construct a final pipeline. We first analyzed the tRNA genes and determined the effect of their overlap with GeneMarkS predicted protein coding genes. We found it is necessary to mask them, as well as other ribosomal RNA genes and pseudogenes before the self training. Secondly, in order to improve the identification of the gene start, we applied the Kullback-Leibler distance to quantify the signal strength of the motif of the ribosomal binding site (RBS) and the distribution of the spacer length upstream to the translation initiation sites. Low GC content genomes have more distinct RBS signal than the high GC content ones. Thirdly, we further explored the possibility to optimize the duration parameter of the hidden Semi-Markov model. By extending the duration, substantial short ORFs were not predicted, leading to an increased specificity while sacrificing less sensitivity. Last but not least, we underwent case studies for extreme low and high GC content genomes. Overall, the result shows that an increase of two percent could be achieved by taking account into these modifications. A manuscript is in preparation to publish these findings.

3.1 Introduction

An upgraded GeneMark.hmm 2.0 was at the core of the GeneMarkS. This new version defined a new hidden state of overlapping genes and corresponding transition probabilities to/from the intergenic regions and the regular isolated genes. This new development enabled the GeneMark.hmm to predict overlapping genes in all possible scenarios: overlap of genes on the same strand such as inside an operon; overlap of genes on opposite strands, head to head, tail to tail and head to tail configurations. Instead of using the RBS signal for start refinement as a post-processing step, the new version integrated it into the Viterbi algorithm. The two component prestart signal consisted of two sub-models: the motif (positional frequency matrix) and the length frequency distribution of the spacer between the ribosomal binding site and translation initiation site.

The first iteration of this iterative training process was to run GeneMark.hmm version 2.0 on the input anonymous DNA sequence, with the coding and non-coding parameters from the heuristic model, to give the first parse. Based on the first parse, the coding and non-coding parameters were derived and the prestart regions were used as input for Gibbs sampling to localize the RBS motif. These *in silico* derived parameters were named pseudonative model and served as the input for all subsequent steps of the regular cycle. This iterative training and refining steps were repeated until a pre-defined convergence point, either 99% identical to the previous run or a certain fluctuation point at a high identity. Interestingly enough, the final output of this iteratively trained pseudonative model could be combined with the heuristic model as a dual-model setup, to effectively model and then predict those typical and atypical genes, respectively.

While the accuracy of gene prediction programs identifying the 3' stop codon has reached 95% and above, there is still room to improve for identification of the other 5' end, the start codon positions. Various programs have used the signals around the start codon to pinpoint translation initiation sites (TIS). These signals include the following features, the length of the regulatory signal, the start codon usage, the operon structure, the spacer length between the ribosome binding site and the start codon, as well as the coding potential of the complete ORF.

Several different approaches exist so far, probabilistic motif and Markov-Bayesian based methods. RBSfinder (Suzek, Ermolaeva et al. 2001) used the 3'-end of the 16S ribosomal RNA as the seed sequence, to which the ribosomal binding site (RBS) sequence binds. The method examines the sequences extending upstream of the candidate start codons, and looks for conserved motifs in these regions. Their score function is based on the number of hydrogen bonds that could be formed between the RBS and the seed sequence. Then the multiple sequences in the training set are aligned and thus a probabilistic model is constructed. Finally, the candidate TISs are scored using this positional weight matrix and the following two rules are considered to determine the output start codon position: 1) the start codons are favored in this descending order: ATG, GTG and then TTG; 2) Given the start codons are the same, the site with the higher scoring RBS is chosen.

Recently, unsupervised procedure was designed to train the parameters of Markov chains to model the TIS regions. TICO (Tech, Pfeifer et al. 2005) and TriTISA (Hu, Zheng et al. 2009) are the representatives of this type. In the vicinity of TIS sites, there are three categories of sequences: the true TIS, false TIS regions upstream and false TIS regions downstream. In terms of sequence evolution, the false TIS upstream are exposed to

neutral evolutions while downstream false TIS exhibits three periodic features. The authors of TriTISA extended their earlier work (Zhu, Hu et al. 2007; Hu, Zheng et al. 2008; Hu, Zheng et al. 2008), and characterized such statistical differences from each category by a non-homogeneous Markov model. As a post-processing tool, TriTISA takes an initial gene TIS prediction as input, and then estimates the transitional positional nucleotide frequencies for each of the three categories, using expectation maximization (EM) scheme. Then, a Bayesian probability that a candidate TIS is a true TIS can be calculated, after taking into account of the prior probabilities. On the other hand, TICO implemented a clustering algorithm to classify the candidate TISs as two categories: *strong* and *weak*. Similarly, the trinucleotide probabilities around the TIS are represented by two classes of inhomogeneous second order Markov models. Then each candidate TIS is scored with a positional weight matrix based on the difference between the log-probabilities of the strong and weak models. In each iteration, the highest positive scored candidate is considered as a strong TIS. Both methods achieved a high accuracy (sensitivity 95.0% and specificity of 99.9%) in a selected experimentally verified gene subset of five genomes, *Aeropyrum pernix*, *Escherichia coli*, *Halobacterium salinarum*, *Natronomonas pharaonis* and *Synechocystis*. Still, the TriTISA authors claimed that TICO's prediction is input sensitive, and it converges within 4-5 steps after undergoing extensive fluctuations. Their results show that TriTISA is extremely robust against the quality of the input, which is of critical importance for a post-processing TIS refinement program.

3.2 Materials

Sequence data of 912 complete prokaryotic genomes were downloaded from the NCBI RefSeq database as of February, 2009. The majority microbial species use the common Genetic code translation table 11 to encode protein (Jukes and Osawa 1993), while 18 *Entomoplasmatales* and *Mycoplasmatales* use table 4 (Bove 1993). In order to build a large enough training set, we didn't include small chromosomes and plasmids shorter than 490 Kbp. The GC content of the genomes under consideration ranges from 16.6% to 74.9%.

3.3 Methods

In the whole GeneMarkS training, we separated into three steps: pre-training processing, model refinement and post-prediction correction.

In the pre-training step, we detect the tRNA and ribosomal RNA genes and mask the tRNA genes.

3.3.1 Deal with long stretch of uncertain nucleotides 'N'

The quality of the input sequence varies from case to case. In RefSeq database, the nucleotides of those complete sequence genomes are all determined, by letter ***A***, ***C***, ***G*** and ***T***. Quite often, the anonymous sequence could derive from low coverage and poorly assembled sequence. Under such circumstances, the input could contain long stretch of uncertain ones, conventionally represented by letter ***N***.

In order to keep the coordinates to report the gene prediction, we substitute long stretch of sequence ***N*** with a gap filler, as illustrated in Figure 3.1. The legend "d" ("r") stands

for direct (reverse) strand, while the “St” (“En”) stands for start (end). This filler is capable of starting/closing a potential ORF on both DNA strands, in any of the six coding frame. This filler need 48 nucleotides to construct. Thus, the number of N’s in the middle is determined by the number of the N’s in the stretch minus 48, in order to keep the original coordinates for the input sequence. Moreover, the number of a stretch of N’s has to reach 48 to be substituted, or it is kept as is.

```
#                    rSt rSt rStdEn dEn dEn     rEn rEn rEndSt dSt dSt
--GAP_FILLER         CATGCATGCATTAACTAACTAANNNNNTTAGTTAGTTAATGCATGCATG
```

Figure 3.1 Gap filler used by GeneMark.hmm for long stretch of 'N' sequences

3.3.2 Masking RNA genes, pseudogenes and tandem repeats

Both of the transfer RNA and ribosomal RNA have different codon usage than the regular protein coding sequences. It is important to evaluate their effect on the model parameter estimation and mask them accordingly.

tRNAScan-SE (Lowe and Eddy 1997) was developed in 1997 and has become the state-of-the-art program for tRNA gene detection. The authors claimed 99% or higher accuracy, about one false positive per 15 GB of genomic DNA sequences after factoring the average length of tRNA genes. The program was run on these 810 microbial genomes to find the tRNA genes.

The best way to find ribosomal RNA genes is still by sequence similarity search. The ribosomal RNA genes include 5S, 16S and 23S rRNA genes. Two well developed databases (Pruesse, Quast et al. 2007; Cole, Wang et al. 2009) are suitable for this purpose.

Moreover, the other two problematic DNA regions for gene finding are the pseudogene and tandem repeats. Even though they usually does not occupy relatively large fraction of microbial genomes, it is a good practice to locate them and mask them out. Researchers from Yale University compiled a resource of pseudogene, namely Pseudofam (Karro, Yan et al. 2007; Lam, Khurana et al. 2009). On the grounds that alanine-rich peptides play role in the local α-helix protein stabilizing (Rohl, Fiori et al. 1999), manual sanity check shall be performed on the repeats detected by program such as Tandem Repeats Finder (Benson 1999). Finally, the regions detected aforementioned should be then masked by letter ***N***.

3.3.3 Refining the model of Ribosomal binding site

Comparing to the NCBI RefSeq annotation (Sayers, Barrett et al. 2009), most state-of-the-art prokaryotic gene finders could readily achieve 90-95% accuracy in terms of 3' stop codon matching, but only average of 80% and lower on 5' start codon. The lack of verified N-terminal verified gene datasets makes the development of accurate start finding even more difficult. Among the few datasets, the EcoGene dataset (Rudd 2000) consisting of 858 gene starts from *E. coli*, is widely used as benchmark. Our program GeneMarkS (Besemer, Lomsadze et al. 2001) correctly detected 856 stop codons and 805 start codons, i.e., 99.8% sensitivity and 93.8% accuracy in terms of start-codon prediction. It is possible that *E. coli* K12 has relatively strong RBS signals and the EcoGene set consists of mostly highly-expressed house-keeping gene. Still, the lack of the golden standard of experiment verified test set exists as a problem for gene finding. GeneMarkS performs relative low in high GC content genomes (Figure 3.2).

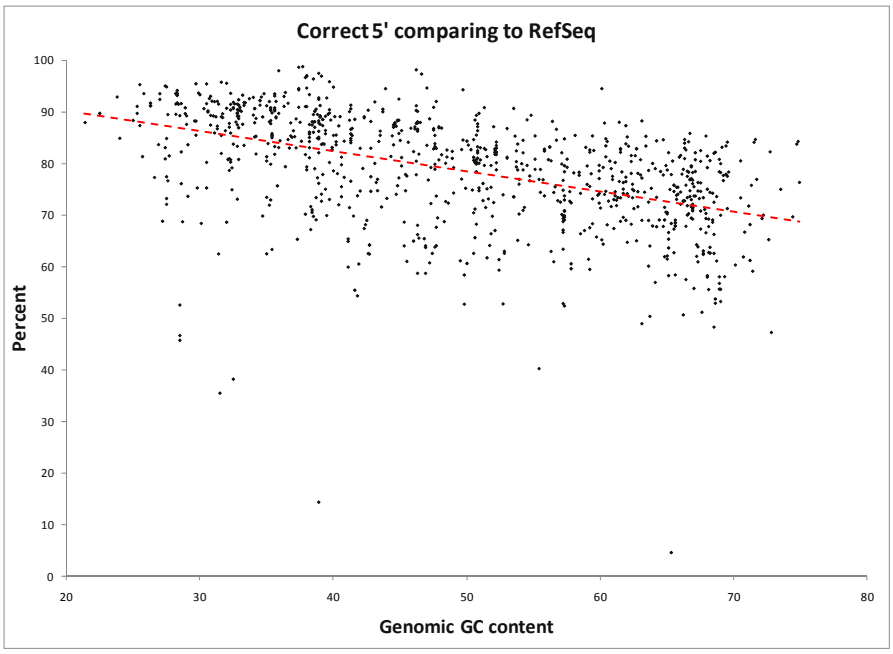


Figure 3.2 Correctly predicted % of gene starts vs. Genomic GC content

The problem of correct start calling can be formalized as follows. For a particular ORF, there could be numerous in-frame ATG's, as well as less frequently used alternative start codons such as GTG or TTG, making the scenario even more complicated. In order to pinpoint the true xTG, GeneMarkS and other methods try to find the signal upstream to the start codon, namely, the RBS site of the Shine-Dalgarno sequence in most microbial genomes. Still, some archaeal genomes lack the SD signal; instead, they use the more eukaryotic-like TATA promoter sequences. In extreme cases, there are no such upstream signals to find, in so-called leaderless mRNAs (Slupska, King et al. 2001). These scenarios need to be considered for gene start prediction.

Gibbs Sampler version 1.0 was used by GeneMarkS to localize the ribosomal binding site in the prestart regions. Section 1.2.8 describes the development history Gibbs sampler programs. The first version, site sampler, was developed in 1993. The motif sampler

allowed 0, 1 and many possible motif models, a parameter specified by the user. The recursive sampler allowed multiple sites per sequence and improved by considering a possible heterogeneous background model. Table 3.1 lists the difference of these three different versions. By default, the GeneMarkS program cuts 26 nucleotide prestart regions to find the RBS site, usually consists of 6 nucleotides. The running time was calculated on a test set of 600 *E. coli* prestart sequences, in order to mimic the routine motif finding task.

Table 3.1 Comparison of several types of Gibbs sampler

Gibbs sampler versions	Year	Running time	Allow multiple or null motif	Allow multiple motif sites per sequence
Site	1993	6 sec	has to be exactly 1	has to be exactly 1
Motif	1995	1 sec	Yes	No
Recursive	2003	11 sec	Yes	Yes

In general, for most prokaryotic genomes, there is only one ribosomal binding site per prestart region. Therefore, the site sampler meets this requirement. In cases of a genome that has majority of leadless mRNAs, the motif sampler shall be used.

Output of Gibbs has two parts, a motif represented as a positional weight matrix (PWM) of 4 nucleotides times 6 positions, and a spacer length distribution indicating the motif location among the possible 21 positions. Owing to the fact that the RBS resides on the noncoding regions, the background nucleotide frequency effect has to be taken into account. Thus, the relative entropy is used, instead of the absolute entropy, in order to calculate the information content of PWM. The Kullback-Leibler divergence is used:

$$D_{KL}(Motif \;||\; Background) = \sum_j \sum_i P_M^j(i) \log_2 \frac{P_M^j(i)}{P_B(i)}, \; (i = \text{A, C, G, T}; j = 1..6)$$

The SD is usually positioned some 5–8 nucleotides upstream from the start codon (Steitz 1975). The optimal spacing depends on exactly which bases at the 3′ end of 16S rRNA (3′-AUUCCUCCAC…5′) participate in the interaction. The spacing requirement can be rationalized by the structural model of AUG binding to the P site of the ribosome. We assumed a uniform distribution as the null background model to calculate the divergence. The frequency is simply the reciprocal of the number of possible motif positions. The number is 21 and it would be transformed to a constant after take logarithm value.

$$D_{KL}(Spacer \,||\, Uniform) = \sum_{j=1}^{21} P_j \log_2 P_j \; .$$

3.3.4 Duration optimization

In the 2001 GeneMarkS paper, the discussion proved that, a combined model, of native and heuristic models, is necessary to catch those atypical genes for higher sensitivity (Table 3.2).

Table 3.2 GeneMarkS training on *E. coli* K12 genome. Accuracy shown for native model and default dual model.

Model	# Annotated	# Predicted	Sn %	Sp %	Average %	# Missed	# Novel
(1) Native	4131	4042	93.7	95.7	94.7	262	173
(2) Native + Heuristic		4389	98.3	92.5	95.4	72	330
(2) - (1)		347	4.6	-3.2	0.7	-190	157

It is a trade-off between missed and novel genes. To recover the 190 missed genes, a comparable price is paid as a set of novel 157 genes. In this case, the average accuracy of sensitivity and specificity gained 0.7%.

One interesting observation is that, significant more predictions (347) were made by adding the heuristic model. Comparing to the number of annotation (4131), a question could be asked: *Is there a way to reduce the false positive rate?*

We have tried different combinations of parameter tuning. For instance, we changed the order of Hidden Markov model. The conclusion was that higher order model could help with marginal gain (data not shown). We also changed the ratios of possible start codons (ATG/GTG/TTG) and did not observe any gain.

Indeed, there is one experiment we have not tried in complete genome, changing the duration of coding and non-coding, which we have reasonable success in metagenomic sequences, as shown in section 2.3.4.

GeneMark.hmm (1998) (Lukashin and Borodovsky 1998) described the Viterbi algorithm as follows:

$$z_1(a_m, d_m) = \max_{(a_1 d_1)\ldots(a_{m-1} d_{m-1})} \left[Prob\{(a_1 d_1)\ldots(a_{m-1} d_{m-1}), b_1 \ldots b_{l-d_m}\} q_{a_{m-1} a_m}\right] p_{a_m(d_m)} p_{a_m(b_{l-d_m+1}\ldots b_l)}$$

Where m is the number of hidden states visted during generation of the first l nucleotides, $q_{a\ m\text{-}1\ a\ m}$ is the probability of transition from hidden state a_{m-1} to state am and $p_{am(dm)}$ is the probabilty of duration d_m for state a_m.

By this way, the original Hidden Markov Model was modified to HMM with duration, to cope with such classification problem, to tell apart the protein coding and noncoding region from anonymous genomic sequences. In the formula above, the probability of $P_{am}(b_{l\text{-}dm+1}\ldots b_l)$ can be calculated by inhomogeneous (coding) and homogeneous

(noncoding) Markov chain probability models. The second term, the coding duration $P_{am}(d_m)$, represents the probability of a stretch of sequence as either coding or noncoding.

Historically, the length distribution probability densities of protein-coding and non-coding regions were derived from the annotated *E.coli* genomic DNA (Lukashin and Borodovsky 1998). The probability density of the coding regions was approximated by γ distribution $g(d) = N_c(d/D_c)^2 exp(–d/D_c)$, where d is the length in nucleotide. On the other hand, the probability density function for the non-coding regions was approximated by exponential distribution $f(d) = N_n exp(–d/D_n)$. Both coefficients N_c and N_n normalizes the distribution function on the interval from 1 to 3000 nt. We call Dc and D_n the coding and non-coding duration. We used the values 150nt and 300nt from *E. coli* genome as the default, shown as the dotted line in Figure 3.3. Skovgaard argued that most microbial genomes were over-predicted, largely due to the spurious short ORFs as false positives (Skovgaard, Jensen et al. 2001). In order to penalize the short ORFs, we extended the durations to be 300 for non-coding and 700 for coding (solid lines).

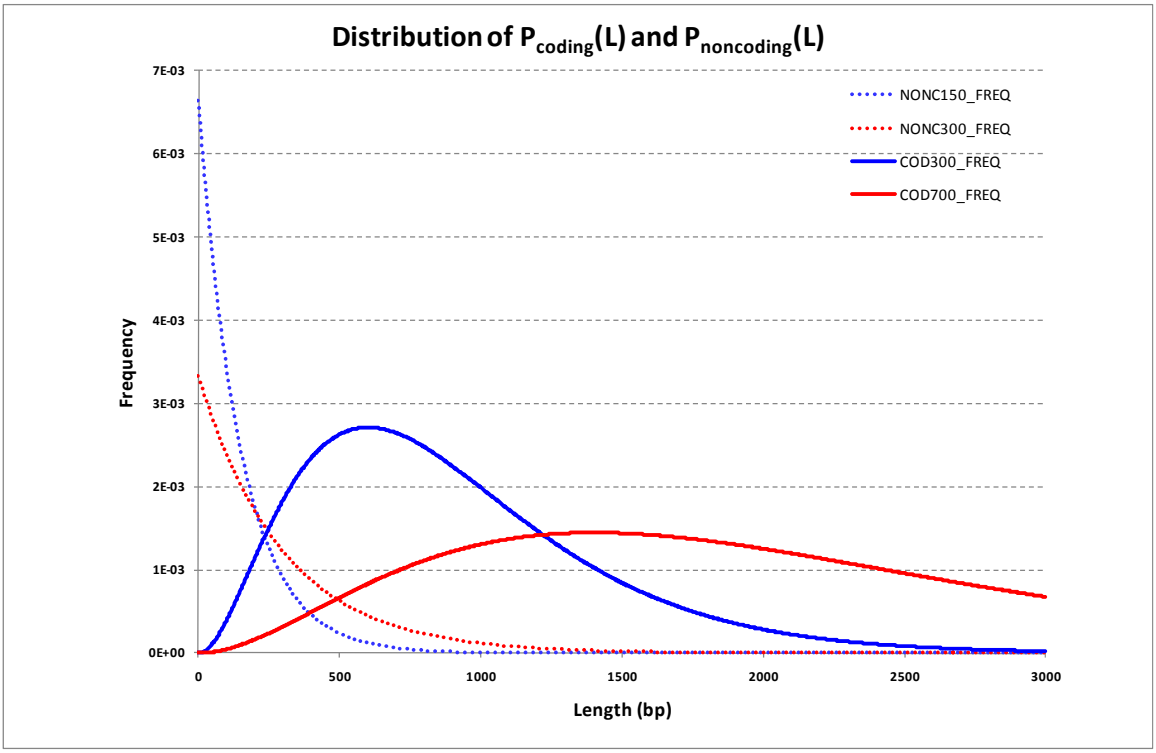


Figure 3.3 Probability density function of non-coding and coding duration.
The default pair (150, 300) and the extended (300, 700) are shown for comparisons. Refer to the text for formula.

In the themes of Baye's and Viterbi theorem, what really matters is the likelihood ratio, namely, the quotient of *P(coding)* divided by *P(noncoding)*, for classification purpose. Figure 3.4 shows the log likelihood ratio versus the ORF size in nt. Dotted and solid lines illustrate the default and the extended durations, respectively.

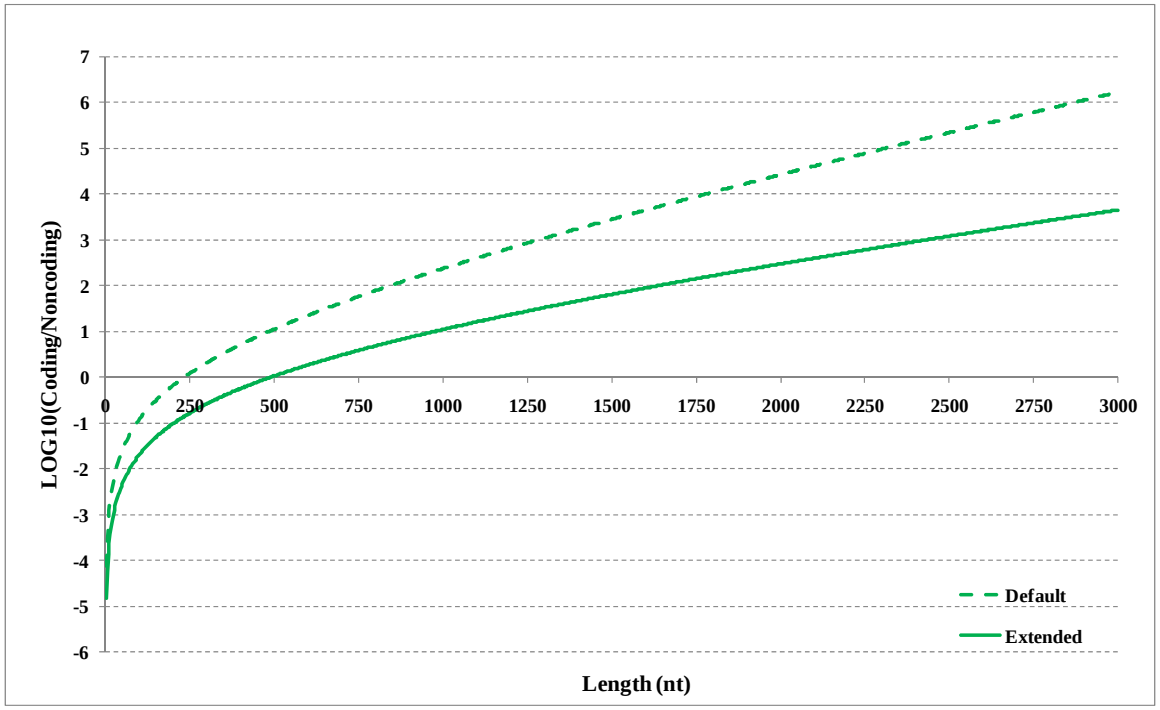


Figure 3.4 Log likelihood ratio of the probability density functions for the coding and noncoding regions.

The two lines intersect with X-axis at 250 and 500 nt. In light of the fact that the logarithm value of 0 corresponds to the ratio of one, this duration extension effectively penalizes those short ORFs less than 500nt instead of 250nt. This change leads to less short gene predicted, thus increases the specificity measure with somewhat loss in sensitivity. Those genuine short coding sequences would be compensated by their native codon usage.

3.4 Results

3.4.1 How many tRNA's fall onto CDS by annotation and prediction

To support efficient protein synthesis, the copy numbers of tRNA were believed to have positive correlation with the expression level of a particular microbial in questions. A positive correlation with GC-content was suggested (Kanaya, Yamada et al. 1999; Zhou, Liu et al. 1999). We repeated the analysis but we didn't observe the correlations between the number of tRNA and the genomic GC content, as illustrated in Figure 3.5.

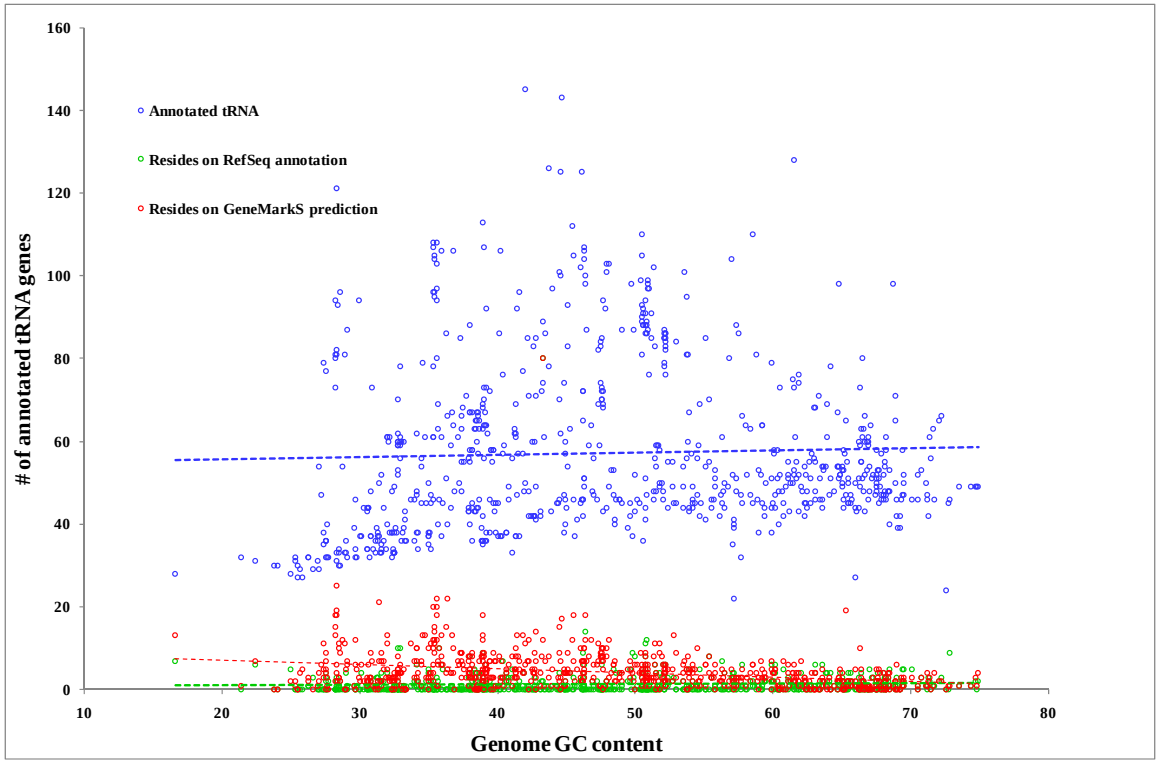


Figure 3.5 Number of tRNA genes versus genomic GC content.

However, we do observe small but significant correlation with the genome size. The tRNA genes are usually GC-rich in order to have the more stable three-hydrogen bonds

for maintaining the 3D cloverleaf structure (Lowe and Eddy 1997). This fact effectively distinguishes them from protein-coding genes and the overlap between these two types of gene is not expected. For the purpose of gene finding, we need to mask these tRNA genes to avoid the false positives. Before that, we have to evaluate the scale of the overlapping problem. We compared the coordinates of tRNA, against the protein-coding genes both from annotation and prediction. The finding is shown in Figure 3.6. On average, of these 809 genomes, there are 57.2 tRNA genes found. 4.2 (1.3) falls onto the protein coding regions by GeneMarkS (RefSeq). In other words, about 4 false positives on average can be avoided by masking tRNA genes prior to the GeneMarkS running.

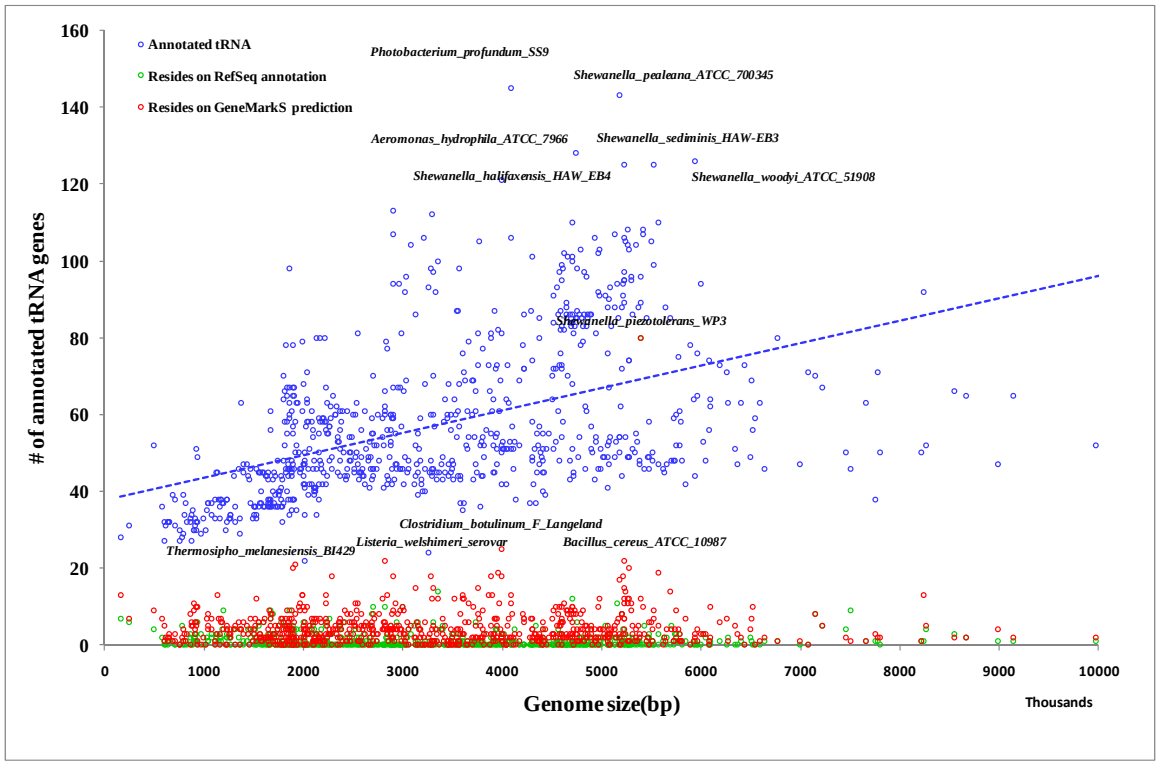


Figure 3.6 The number of tRNA genes versus the genome size.
Top 5 genomes with most tRNA genes are labeled. The top five genomes with most tRNA residing on GeneMarkS predictions are also labeled.

3.4.2 Refine prestart regions

3.4.2.1 Prestart region analysis

To pinpoint the ribosomal binding site, we try to find the ribosomal binding sites signal hidden in the prestart regions. In order to achieve this goal, we used the prestart regions determined by the gene predictions called by GeneMarkS native model. Two items were checked: 1) The distance to the upstream gene and 2) the distance to the first downstream potential start codon (could be ATG/GTG/TTG). The joint distribution of the distances is illustrated in Figure 3.7 of the 4,044 genes of the *E. coli* K12 genome.

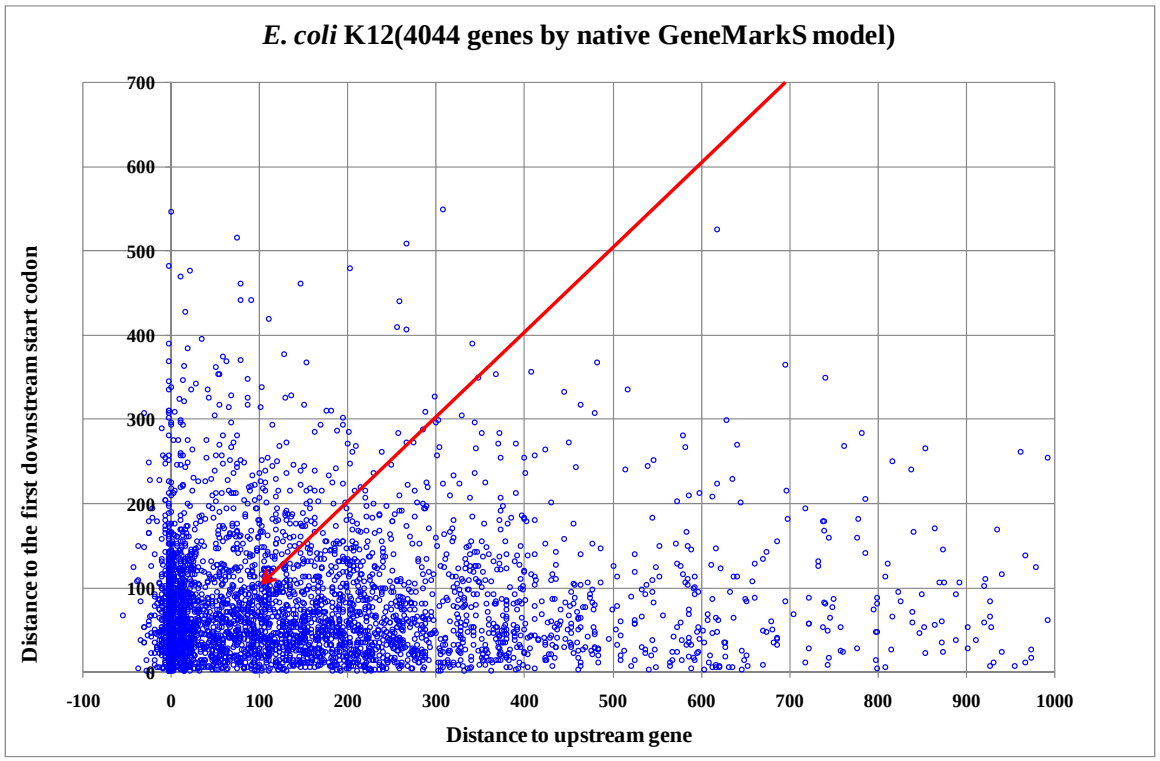


Figure 3.7 The joint distribution of distance to the upstream gene and the first potential downstream start codon

Two observations could be made. 1) Those genes to the left of Y-axis have negative distance value to the upstream gene. In other words, they overlap with the upstream gene.

2) Large proportion of genes have small distance (<50nt) to the previous genes. This fact corresponds to the operon structure in most microbial genomes (Pertea, Ayanbule et al. 2009).

We assume that the best candidates for RBS detection are those genes with long distances to both upstream and the potential downstream start codon. They are more likely to be the first gene in a particular operon. These candidates reside on the top right corner of the joint distribution plot. Along the 45 degree line as the red arrow shows, the prestart regions show a descending order of quality given that our assumption is correct.

3.4.2.2 Sequence logo for six genomes

Our collaborator at PKU compiled a dataset of five genomes, namely, *Aeropyrum pernix, Halobacterium salinarum, Natronomonas pharaonis and Synechocystis.* Each species has a subset of genes whose 5' start positions were verified by wet-lab experiment. We observed similar pattern of joint prestart distance distribution in these four species as well (data not shown). Moreover, the well-annotated *Bacillus subtilis* was added to make a six species test set. The consensus sequences were found by running GeneMarkS and Gibbs Sampler. Figure 3.8 shows the Ribosomal Binding Site (RBS) sequence logos and the spacer distributions of these 6 genomes.

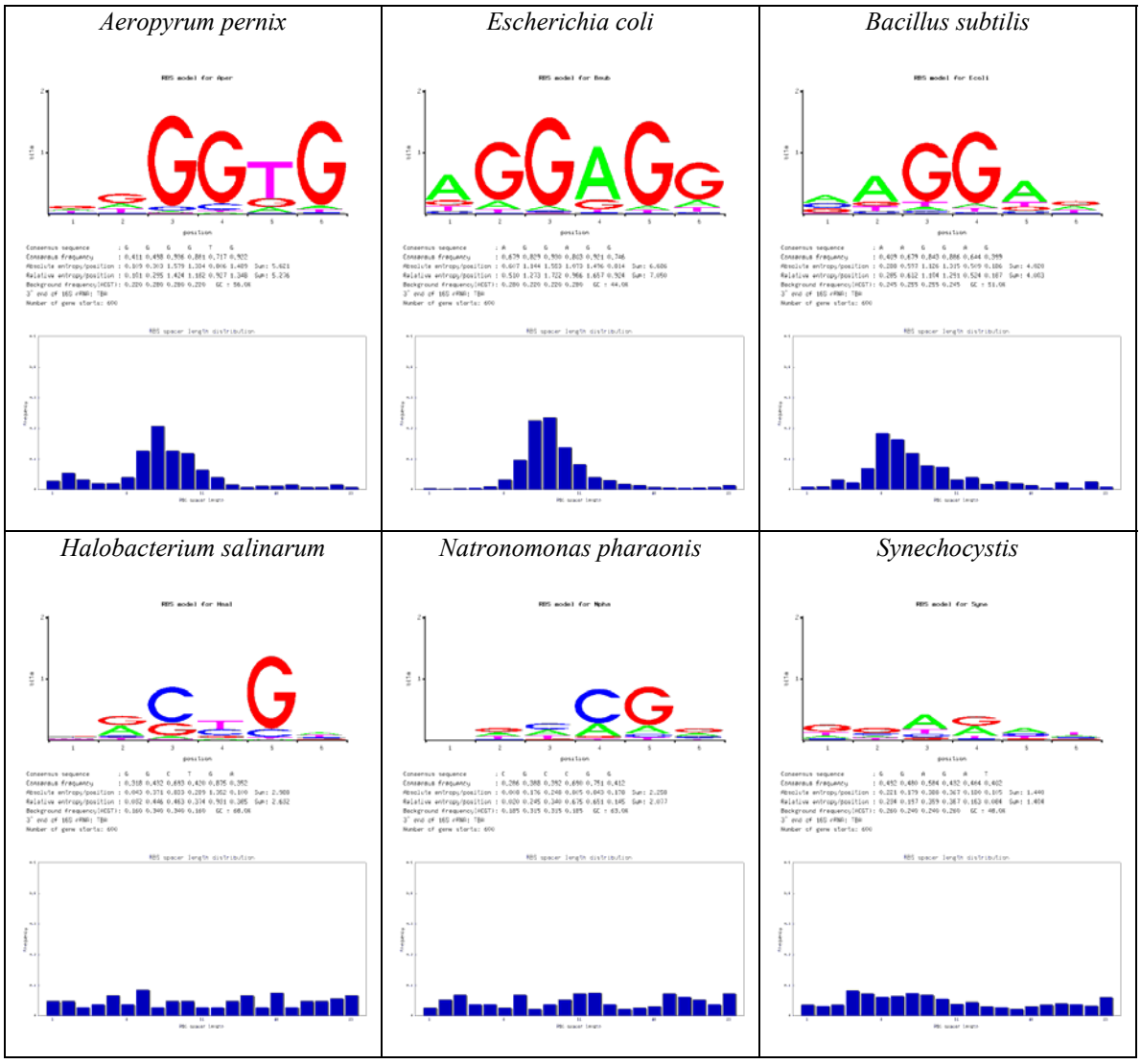


Figure 3.8 Sequence logo and spacer distribution of ribosomal binding site for six genomes.

The G+C content varies among the genomes, so it is necessary to consider the ATCG composition of the noncoding regions as the background. In this way, the relative information content is a true measure for the RBS site. The top three genomes have well conserved RBS site and corresponding strong localization signal of the spacer length distribution. The bottom three genomes show a uniform pattern of the spacer distribution and weak sequence logo signal.

3.4.2.3 Joint distribution of information content for prestart region in microbial genomes

We did an analysis on the prestart regions of 810 microbial genomes. We categorized all genomic sequences into 3 bins based on genomic GC-content, low (40% and less), medium (40% to 60%) and high (60% and more). In a similar fashion as Figure 3.8, we derived the positional weight matrix and the spacer length distribution. The Kullback-Leibler divergences to the background model were calculated using the formula defined in section 3.3.3. This divergence could also be called the relative entropy of information content.

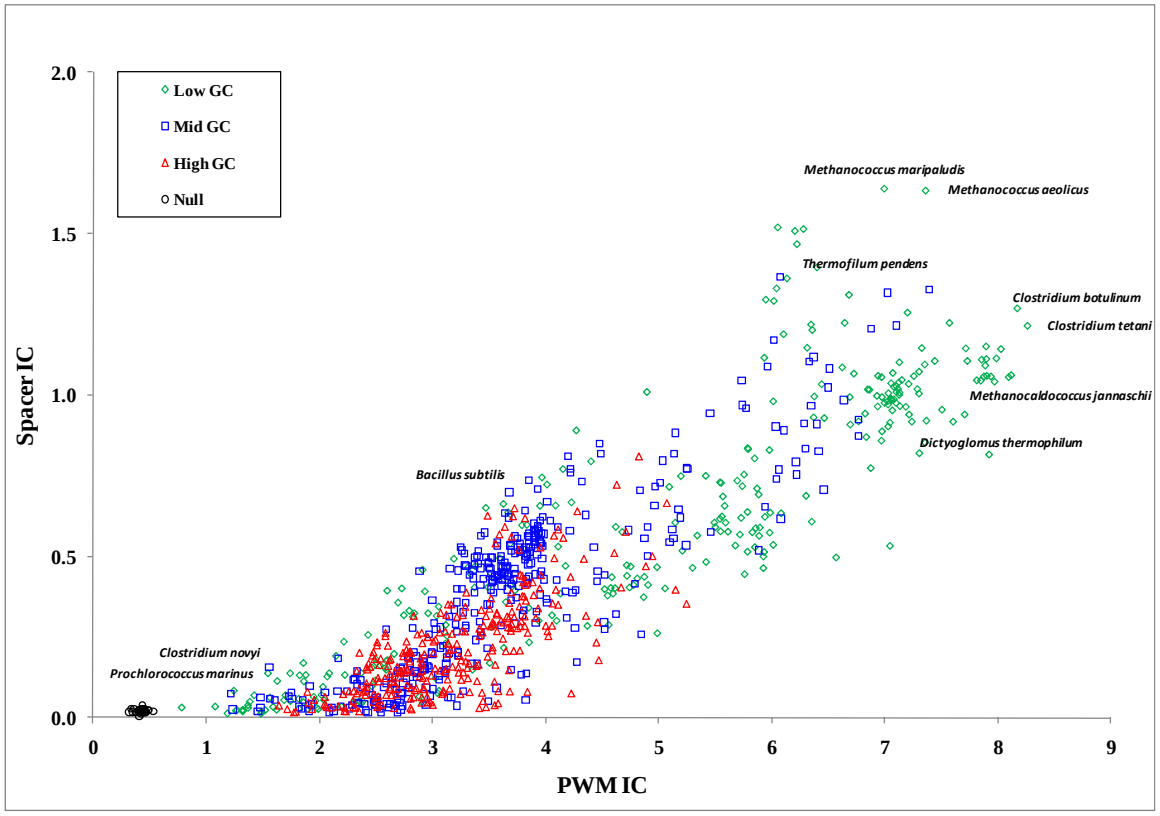


Figure 3.9 Joint distribution of information content for prestart region in microbial genomes.
Low, middle and high GC-content genomes are represented in colors of green, blue and red, respectively.

Figure 3.9 clearly shows that a strong linear positive correlation does exist between the information content of the positional weight matrix and the spacer localization. Several extreme low/high value genomes were labeled. The ribosomal binding site is usually AGGAG-like Shine-Dalgarno sequence, which pairs with the 3'end of the 16S ribosomal RNA during translation. Such G-rich sequences tend to stand out in low-GC genomes. On the other hand, in high GC genomes, these signals are not so obvious. Figure 3.2 in the methods section showed a dropped accuracy for gene start in high GC content genomes. This could be the result of the high GC genomes have relative low information content in PWM and spacer distributions (red dots clustered at the bottom-left corner comparing to green dots clustered at the upper-right corner in Figure 3.9)

Those black dots to the far left bottom end are simulations generated by a null distribution. Three sets of sequences are generated based on multinomial A,C,G,T model using GC equals to 30, 50 and 70 percent, assuming A=T, C=G. Each set composes of 600 sequences and was used as input for Gibbs program. A clear separation could be observed of the real genome data points from these null dots. This experiment proved that the ribosomal binding site could be used to refine the 5' calling.

We also tried to find the relationship of start codon prediction accuracy against the RBS total information content calculated as: $\sqrt{IC_{PWM}^2 + IC_{Spacer}^2}$, the positive correlation can be observed in Figure 3.10 .

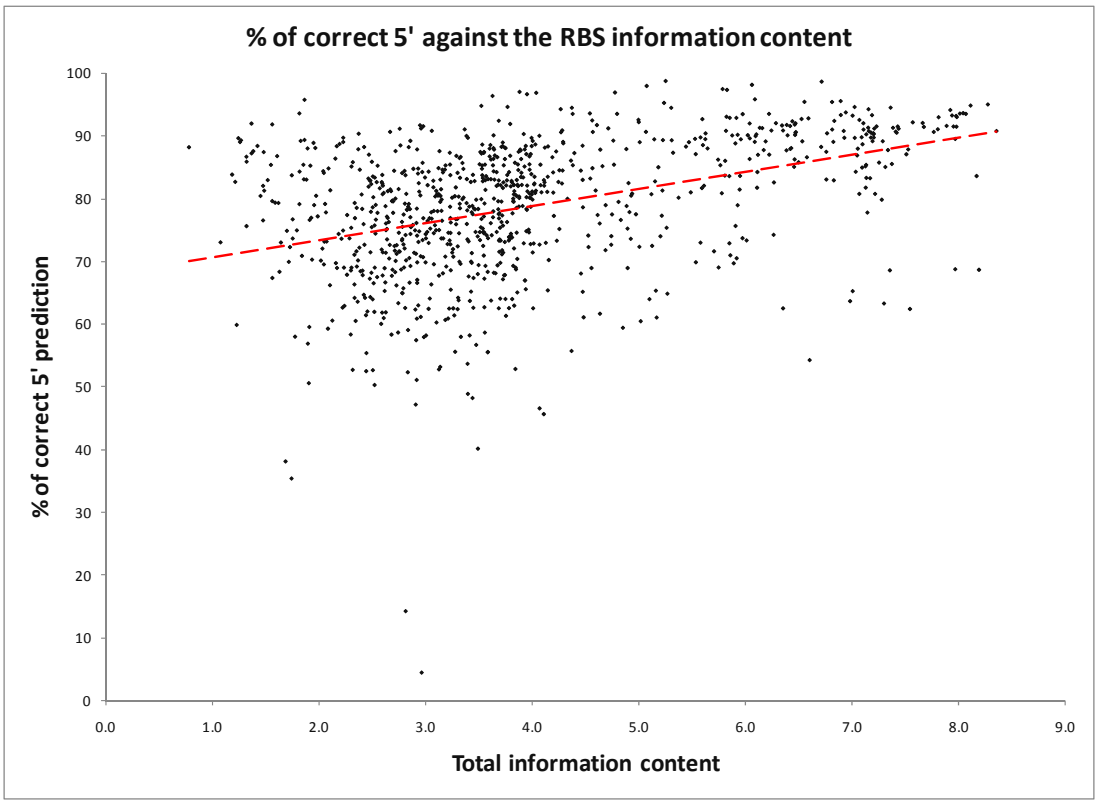


Figure 3.10 Percent of correct 5' start predicted against the genomic RBS information content.

Another issue is the start codon preference of the xTG (ATG/GTG/TTG) among different species. We used the final iteration prediction of GeneMarkS, to calculate the proportion of xTG, result shown in Figure 3.11. The reason we didn't use the RefSeq annotation for this test is that, the proportion is from the convergence point of the self-training process and is more objective. ATG is the mostly frequently used start codon of microbial genomes. It is a GC-less codon comparing to GTG. We did observe a preference of GTG in high GC genomes.

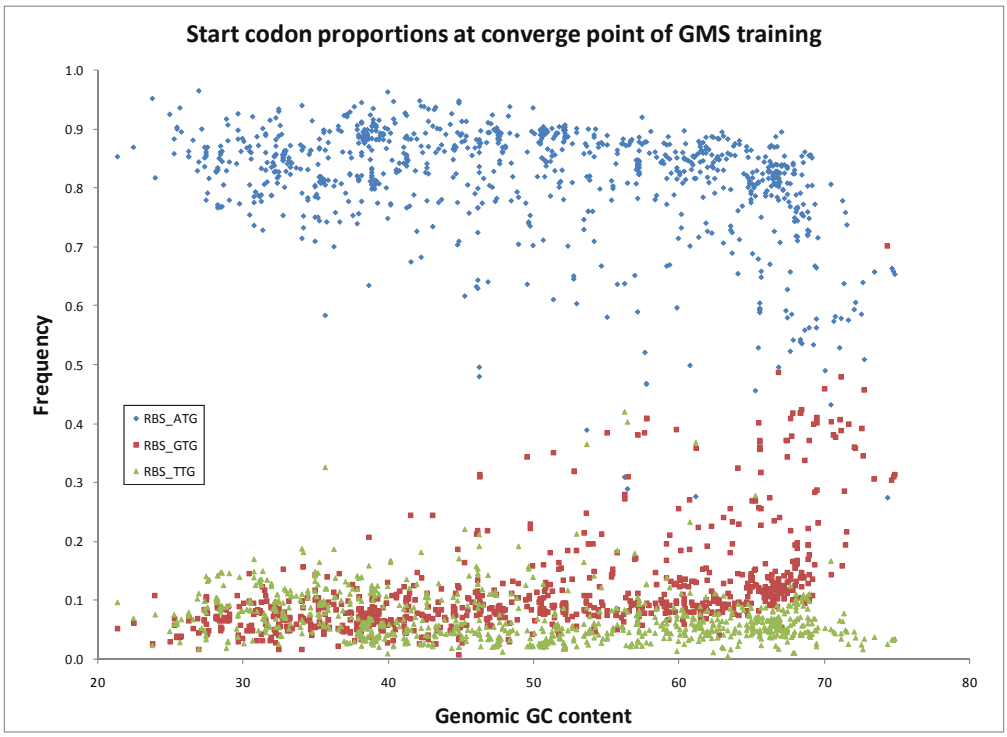


Figure 3.11 xTG composition at the convergence point of GeneMarkS in 810 genomes.

3.4.3 Genomes case by case

3.4.3.1 Low information content genomes

We specify the low boundary of combined information content (IC) to be 0.2 and there are 325 genomes. The GC content and size ranges for these 325 genomes are [27.4%, 74.4%] and [490K, 13M], respectively, both cover the full spectrum of the data set we have in hand. Even though we observe a positive correlation for these low IC genomes, we wanted to see the effect of switching off the RBS model of GeneMarkS training. We used the RBS on mode as the base line. The difference of accuracy is plotted against genomic GC, in three categories: sensitivity of 3' (blue), specificity of 3' (green), and the exact of 5' for those 3' predicted correctly (red, the major concern) as in Figure 3.12.

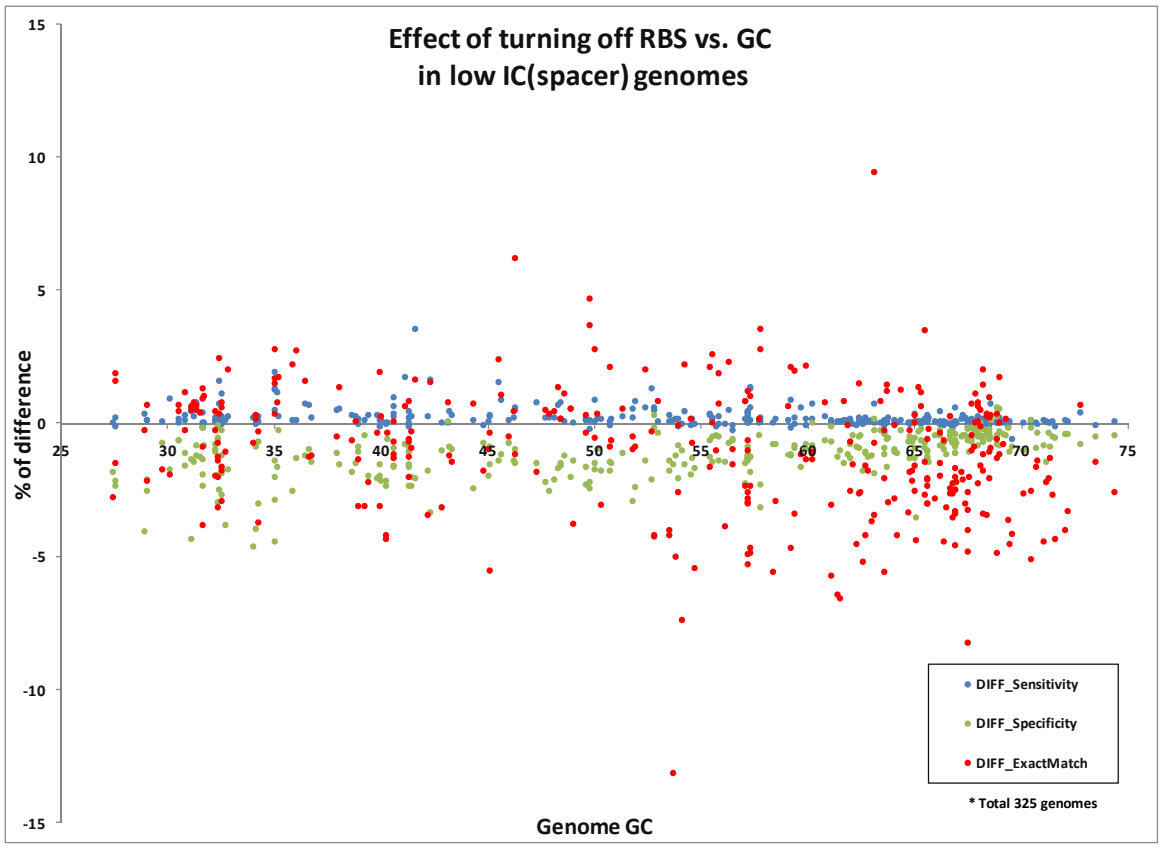


Figure 3.12 Accuracy change when turning off RBS in low IC(RBS) genomes

The difference of sensitivity (blue) is greater than 0 in all genomes, due to the fact that more genes are predicted, leading to more stops at 3' are found correctly, but paying a price of the lower specificity (green). Contrary to our expectation, the exact percentage drops in more genomes and especially for the extreme high GC content genomes. In general, there are two types of RBS-lack genomes, TATA-box like and leaderless mRNA. We can extend the prestart region up to 50nt to localize the TATA-like box signal. On the other hand, we have to turn off the RBS module for those leadless mRNA genomes, such as *Pyrobaculum aerophilum* (Slupska, King et al. 2001).

3.4.3.2 Switch off heuristic model for genomes with G+C content <30%.

The smallest genome sequenced so far is *Candidatus Carsonella ruddii PV* (Nakabachi, Yamashita et al. 2006) with GC content = 16.5% and genome size of 159,662 bp. Keep in mind that we are mostly interested in the large chromosomes, for the fact that the small genomes/plasmids would have small number of annotated genes, leading to a larger variation when calculating sensitivity and specificity. There are total 60 genomic sequences falling into the low GC region and their sizes range from 601,943 to 6,000,632bp.

The default GeneMarkS output mode combines the last iteration of native model from training and the heuristic 1999 HAL model corresponding to the genomic GC. For low GC (<30%) genomes, the combined model uses native model as COD1 (coding 1), heuristic model as COD2 (coding 2) and heuristic non-coding as NONC (non-coding). We turned off the coding 2 model and substituted the non-coding model to be the native one derived from GeneMarkS training, and the comparison is listed in Table 3.3.

Table 3.3 Model settings of turning off heuristic.

	Before (default GeneMarkS)	**After (Turn off heurisic)**
Coding 1	Native (GC)	Native (GC)
Coding 2	Heuristic 30% model	off
Non-coding	Heuristic 30% model	Native (GC)
RBS	Native RBS	Native RBS

Figure 3.13 Accuracy difference after turning off heuristic model in low GC-content genomes.Figure 3.13 shows the difference of the accuracy. The prediction is compared to RefSeq annotation, in terms of 3' stop codon accuracy of sensitivity (blue) and specificity (green), as well as the sum (red) of these two values, against "genomic GC content" and "# of annotated genes" (which is also positively correlated to genome size).

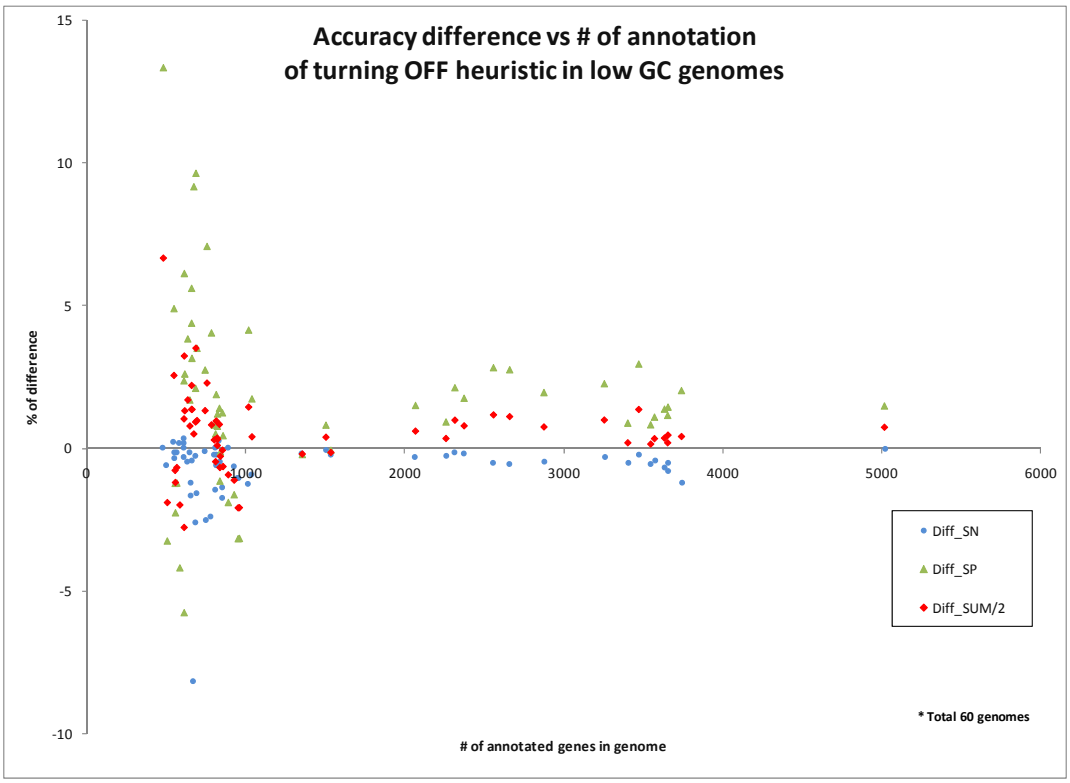


Figure 3.13 Accuracy difference after turning off heuristic model in low GC-content genomes.

In experience, heuristic model was designed to predict atypical genes, which are commonly relatively short genes. As a result (Table 3.4), in most genomes, the sensitivity benchmark drops down as the specificity increases even more, leading to the increase of average of sensitivity plus specificity. The most striking outlier is genome *Aster yellows witches-broom phytoplasma AYWB*, showing a minus 8.19 and a plus 9.16 in Sn and Sp, respectively. Interestingly, the last column, the normalized % of number of predictions has a significant effect on Sn/Sp pairs. For instance, there are 671 *Aster* genes annotated in RefSeq. The default GeneMarkS program predicted 817 genes, leading to 160 new predictions.

Table 3.4 Accuray in low GC-content genomes.
Table is sorted in ascending order of specificity measure.

Accession #	Species	GC (%)	size (bp)	Genetic code	anno #	Kingdom	Δ Sn (%)	Δ Sp (%)	Δ (Sn+Sp)/2 (%)	normalized % Δ of predictions
NC_007292	*Candidatus_Blochmannia_pennsylvanicus_B*	29.6	791654	11	610	Bacteria	0.2	-5.8	-2.8	6.6
NC_005061	*Candidatus_Blochmannia_floridanus*	27.4	705557	11	583	Bacteria	0.2	-4.2	-2.0	5.0
NC_004545	*Buchnera_aphidicola*	25.3	615980	11	504	Bacteria	-0.6	-3.3	-1.9	3.4
NC_006831	*Ehrlichia_ruminantium_Gardel*	27.5	1499920	11	950	Bacteria	-1.1	-3.2	-2.1	2.4
NC_006832	*Ehrlichia_ruminantium_str._Welgevonden_C*	27.5	1512977	11	958	Bacteria	-1.1	-3.2	-2.1	2.4
NC_011833	*Buchnera_aphidicola_5A__Acyrthosiphon_p*	26.3	642122	11	555	Bacteria	-0.2	-2.3	-1.2	2.5
NC_005295	*Ehrlichia_ruminantium_Welgevonden_UPSA*	27.5	1516355	11	888	Bacteria	0.0	-1.9	-1.0	2.4
NC_007354	*Ehrlichia_canis_Jake*	29.0	1315030	11	925	Bacteria	-0.6	-1.6	-1.1	1.2
NC_011834	*Buchnera_aphidicola_Tuc7__Acyrthosiphon_*	26.3	641895	11	553	Bacteria	-0.4	-1.2	-0.8	1.1
NC_002528	*Buchnera_sp*	26.3	640681	11	564	Bacteria	-0.2	-1.2	-0.7	1.2
NC_000963	*Rickettsia_prowazekii*	29.0	1111523	11	835	Bacteria	-0.2	-1.2	-0.7	1.1
NC_007205	*Candidatus_Pelagibacter_ubique_HTCC1062*	29.7	1308759	11	1354	Bacteria	-0.2	-0.2	-0.2	0.0
NC_006142	*Rickettsia_typhi_wilmington*	28.9	1111496	11	838	Bacteria	-0.5	-0.1	-0.3	-0.4
NC_007681	*Methanosphaera_stadtmanae*	27.6	1767403	11	1534	Archaea	-0.3	-0.1	-0.2	-0.2
NC_008277	*Borrelia_afzelii_PKo*	28.3	905394	11	855	Bacteria	-1.8	0.4	-0.7	-2.2
NC_011728	*Borrelia_burgdorferi_ZS7*	28.5	906707	11	808	Bacteria	-1.5	0.5	-0.5	-2.1
NC_010673	*Borrelia_hermsii_DAH*	29.8	922307	11	819	Bacteria	-0.6	0.8	0.1	-1.5
NC_011244	*Borrelia_recurrentis_A1*	27.5	930981	11	800	Bacteria	-0.3	0.8	0.3	-1.3
NC_012039	*Campylobacter_lari_RM2100*	29.7	1525460	11	1503	Bacteria	-0.1	0.8	0.4	-0.9
NC_009697	*Clostridium_botulinum_A_ATCC_19397*	28.2	3863450	11	3548	Bacteria	-0.6	0.8	0.1	-1.5
NC_009698	*Clostridium_botulinum_A_Hall*	28.2	3760560	11	3404	Bacteria	-0.5	0.9	0.2	-1.5
NC_008710	*Borrelia_turicatae_91E135*	29.1	917330	11	818	Bacteria	-0.2	0.9	0.3	-1.2
NC_009850	*Arcobacter_butzleri_RM4018*	27.0	2341251	11	2259	Bacteria	-0.3	0.9	0.3	-1.2
NC_009495	*Clostridium_botulinum_A*	28.2	3886916	11	3572	Bacteria	-0.5	1.1	0.3	-1.7
NC_010520	*Clostridium_botulinum_A3_Loch_Maree*	28.3	3992906	11	3655	Bacteria	-0.8	1.1	0.2	-2.1
NC_011229	*Borrelia_duttonii_Ly*	27.6	931674	11	820	Bacteria	-0.6	1.2	0.3	-2.0
NC_001318	*Borrelia_burgdorferi*	28.6	910724	11	851	Bacteria	-1.4	1.2	-0.1	-2.7
NC_009699	*Clostridium_botulinum_F_Langeland*	28.3	3995387	11	3635	Bacteria	-0.7	1.4	0.3	-2.3
NC_006156	*Borrelia_garinii_PBi*	28.3	904246	11	832	Bacteria	0.2	1.4	0.8	-1.2
NC_010516	*Clostridium_botulinum_B1_Okra*	28.3	3958233	11	3657	Bacteria	-0.5	1.4	0.4	-2.1
NC_009617	*Clostridium_beijerinckii_NCIMB_8052*	29.9	6000632	11	5020	Bacteria	0.0	1.5	0.7	-1.8
NC_003454	*Fusobacterium_nucleatum*	27.2	2174500	11	2067	Bacteria	-0.3	1.5	0.6	-1.9
NC_011374	*Ureaplasma_urealyticum_serovar_10_ATCC_*	25.8	874478	4	646	Bacteria	-0.2	1.7	0.8	-2.0
NC_004432	*Mycoplasma_penetrans*	25.7	1358633	4	1037	Bacteria	-1.0	1.7	0.4	-2.8
NC_004557	*Clostridium_tetani_E88*	28.7	2799251	11	2373	Bacteria	-0.2	1.8	0.8	-2.7
NC_007633	*Mycoplasma_capricolum_ATCC_27343*	23.8	1010023	4	812	Bacteria	0.0	1.9	0.9	-2.1
NC_008261	*Clostridium_perfringens_ATCC_13124*	28.4	3256683	11	2876	Bacteria	-0.5	1.9	0.7	-2.6
NC_009089	*Clostridium_difficile_630*	29.1	4290252	11	3742	Bacteria	-1.2	2.0	0.4	-3.6
NC_006055	*Mesoplasma_florum_L1*	27.0	793224	4	682	Bacteria	-0.3	2.1	0.9	-2.5
NC_008593	*Clostridium_novyi_NT*	28.9	2547720	11	2315	Bacteria	-0.2	2.1	1.0	-2.4
NC_010723	*Clostridium_botulinum_E3_Alaska_E43*	27.4	3659644	11	3256	Bacteria	-0.3	2.3	1.0	-2.7
NC_010503	*Ureaplasma_parvum_serovar_3_ATCC_2781*	25.5	751679	4	609	Bacteria	-0.3	2.4	1.0	-2.8
NC_002162	*Ureaplasma_urealyticum*	25.5	751719	4	614	Bacteria	0.0	2.6	1.3	-2.8
NC_009497	*Mycoplasma_agalactiae_PG2*	29.7	877438	4	742	Bacteria	-0.1	2.7	1.3	-3.4
NC_003366	*Clostridium_perfringens*	28.6	3031430	11	2660	Bacteria	-0.6	2.8	1.1	-3.5
NC_008262	*Clostridium_perfringens_SM101*	28.2	2897393	11	2558	Bacteria	-0.5	2.8	1.2	-3.6
NC_010674	*Clostridium_botulinum_B_Eklund_17B*	27.5	3800327	11	3473	Bacteria	-0.3	2.9	1.3	-3.4
NC_007294	*Mycoplasma_synoviae_53*	28.5	799476	4	659	Bacteria	-0.5	3.2	1.3	-4.4
NC_006360	*Mycoplasma_hyopneumoniae_232*	28.6	892758	4	691	Bacteria	-1.6	3.5	1.0	-5.8
NC_006908	*Mycoplasma_mobile_163K*	25.0	777079	4	633	Bacteria	-0.5	3.8	1.7	-4.9
NC_002771	*Mycoplasma_pulmonis*	26.6	963879	4	782	Bacteria	-2.4	4.0	0.8	-6.9
NC_005364	*Mycoplasma_mycoides*	24.0	1211703	4	1016	Bacteria	-1.3	4.1	1.4	-7.3
NC_007332	*Mycoplasma_hyopneumoniae_7448*	28.5	920079	4	657	Bacteria	-1.7	4.4	1.3	-7.8
NC_004061	*Buchnera_aphidicola_Sg*	25.3	641454	11	546	Bacteria	0.2	4.9	2.5	-6.4
NC_007295	*Mycoplasma_hyopneumoniae_J*	28.5	897405	4	657	Bacteria	-1.2	5.6	2.2	-8.7
NC_004344	*Wigglesworthia_brevipalpis*	22.5	697724	11	611	Bacteria	0.3	6.1	3.2	-7.2
NC_005303	*Onion_yellows_phytoplasma*	27.7	860631	11	754	Bacteria	-2.5	7.1	2.3	-17.1
NC_007716	*Aster_yellows_witches-broom_phytoplasma*	26.9	706569	11	671	Bacteria	-8.2	9.2	0.5	-21.6
NC_010544	*Candidatus_Phytoplasma_australiense*	27.4	879959	11	684	Bacteria	-2.6	9.6	3.5	-23.0
NC_011047	*Candidatus_Phytoplasma_mali*	21.4	601943	11	479	Bacteria	0.0	13.3	6.7	-18.4

3.4.3.3 Shift heuristic model to lower G+C content for genomes with G+C content > 65%

Similarly to low GC cases, we performed trail analysis on the subset of genomic sequences with GC content more than 65%, a total 173 genomes. Modifying one item at a time, the following three new settings were tried out. The changes are indicated by bold font in Table 3.5.

Table 3.5 Model settings for high GC content genomes.

	Default	Setting A	Setting B	Setting C
Coding 1	Native (GC)	Native (GC)	Native (GC)	Native (GC)
Coding 2	Heuristic (GC)	**Heuristic (GC-10)**	Heuristic (GC)	**Heuristic (GC-10)**
Non-coding	Heuristic (GC)	Native (GC)	**Native (GC)**	**Native (GC)**

Figure 3.14 illustrates the average of sensitivity plus specificity, against genome size for the purpose of separation.

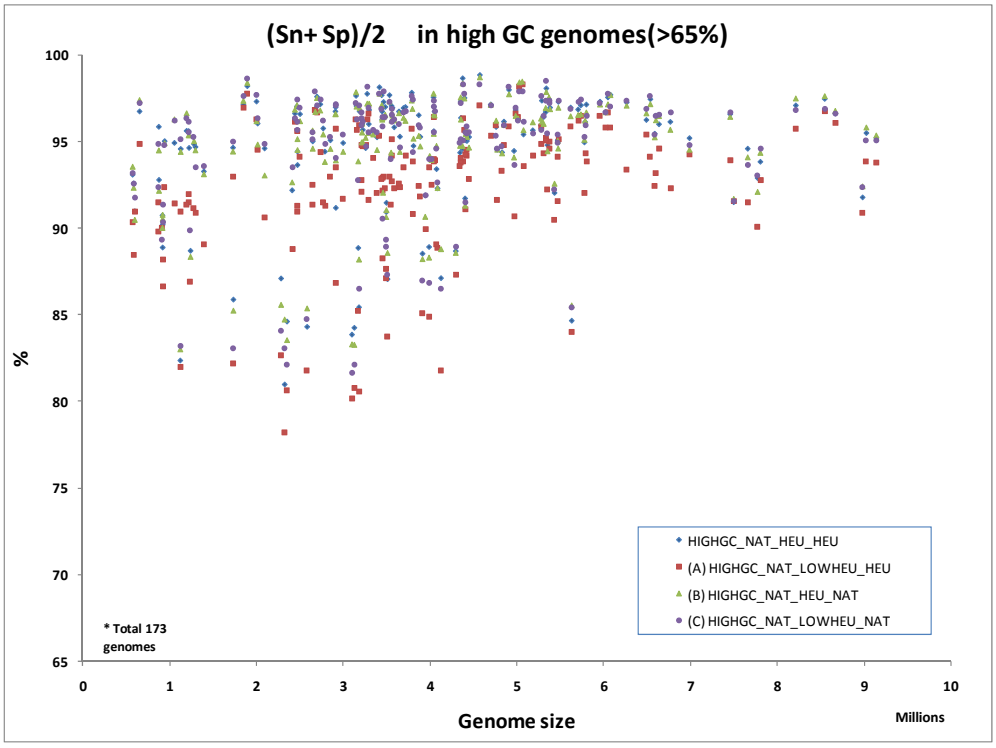


Figure 3.14 Average accuracy for high GC content (>65%) genomes, under four different settings.

After comparing the sensitivity and specificity, we can observe the following.

1) Setting A, use a lower (10%) GC content COD2 model.

   The sensitivity increased across almost all genomes while the specificity drops down significantly. The calculated average indicated this shifting process is not worthwhile.

2) Setting B, use the native non-coding model from the GeneMarkS training.

   When the noncoding model was changed to native, the specificity is always the top among these four combinations, as well as the average accuracy measure.

3) Setting C, make both changes.

   This accuracy of setting C is a combinatorial effect of A and B. However, the average accuracy is similar to B.

4) Again, it boils down to the observation of # of genes predicted. The following observation is common across all high GC species:

   The number of genes predicted: (A) > Default > (C) > (B) (Figure 3.15)

A further analysis on how to control the number of genes predicted is much needed.

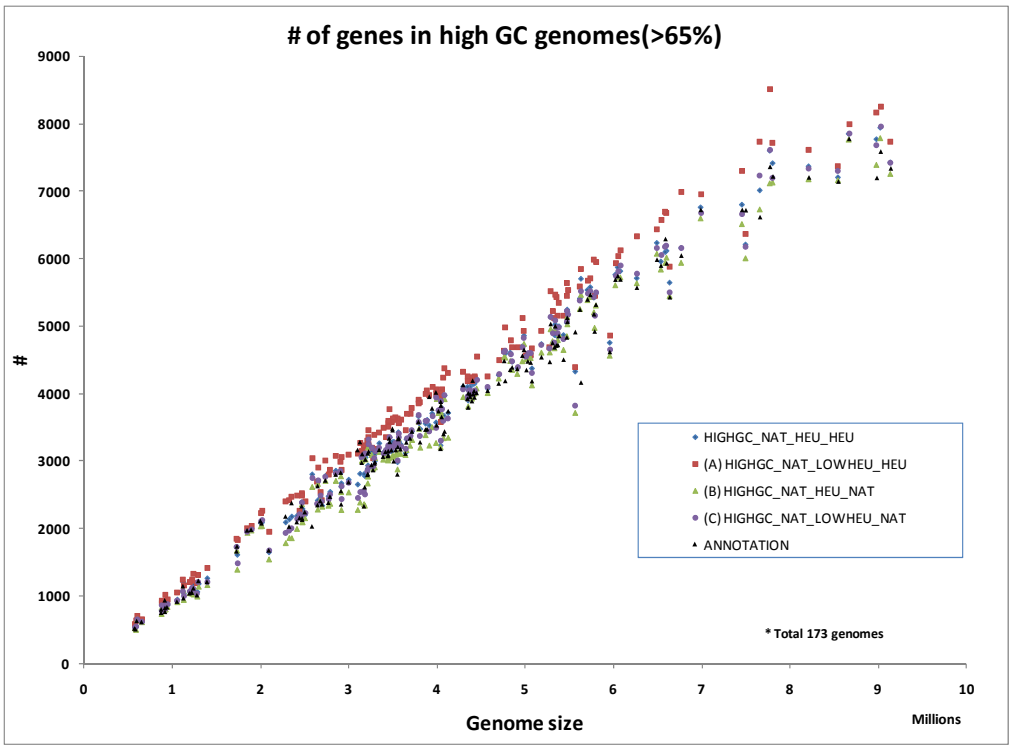


Figure 3.15 Number of genes predicted by four settings of model.

3.4.4 The stability of Gibbs Sampler

GeneMarkS employs an iterative Expectation Maximization (EM) fashion for parameter approximation. At the end of each iteration, the program checks the difference between current run and the previous run. The default convergence criterion is set to be either 99% exactly the same or 10 iterations, whichever comes first.

The users have given us some feedback about the stability issue: They observe GeneMarkS output varies from runs. In *Bacillus subtilis* genome, there is a variation of about 20 and 50 genes in 3' and 5' gene locations, respectively.

This is largely due to the motif finding program, Gibbs Sampler (Lawrence, Altschul et al. 1993), applies a non-deterministic approach to localized the ribosomal binding sites. The

first version of Gibbs sampler could not handle large numbers of input data. In the current GeneMarkS implementation, we empirically derive the RBS signal from 600 prestart regions, as the input of Gibbs Sampler version 1.0. A new 3.1 version of Gibbs Sampler is available, released in August 2009. The new features that could be used by our group are: i) can handle more input sequences; ii) a pseudo-count weight that could reduce the variation among sampling runs; iii) optional more iterations and iv) a Maximum *a posterior* (MAP) alignment.

It is possible to improve by doing the following.

1) Take more prestart regions as input for Gibbs Sampler;
2) Select better quality prestart regions, which have a certain distance from both the upstream gene and the first potential downstream start codon;
3) Exclude those prestart regions overlapping the upstream genes;
4) Upgrade to Gibbs Sampler version 3.0, in order to use the MAP alignment, instead of the near optimal alignment.

The following table (Table 3.6) shows the identity differences between the n-th iteration and the previous n-1 th iteration of GeneMarkS training. These three genomes have high information content RBS site signal. Therefore, they are good candidates for such test. Note that the new Gibbs version 3.1 reached the identity difference level at about 99.5%. In terms of gene counts, the *E. coli* genome has 4100 genes annotated, a 99.5% identity would translate to about 20 genes or so (4100*0.5), a reduce of one-fold from the variation of 50 genes by Gibbs version 1.0.

Table 3.6 The difference in percentage between two successive iteration of GeneMarkS training.
The 99% identity convergence checkpoint was turned off, in order to keep the iteration going till the 10th.

	Aeropyrum		*Escherichia coli*		*Bacillus subtilis*	
# of iterations	**Gibbs ver 1.0**	**Gibbs ver 3.1**	**Gibbs ver 1.0**	**Gibbs ver 3.1**	**Gibbs ver 1.0**	**Gibbs ver 3.1**
1	61.33	59.49	71.94	70.73	67.53	66.77
2	85.62	85.96	95.04	96.17	94.87	95.74
3	95.22	91.36	97.91	98.59	97.42	99.24
4	95.57	99.14	97.89	99.27	98.57	99.52
5	97.77	99.55	97.65	99.05	98.93	99.27
6	97.86	96.95	97.43	99.42	98.50	99.70
7	98.33	99.04	98.37	99.38	98.35	99.70
8	97.47	96.92	98.02	99.33	97.88	99.50
9	95.95	99.43	97.54	99.48	98.48	99.59
10	95.36	99.61	97.49	99.33	98.21	99.67

3.4.5 Duration test

Similar to the methods introduced in section 2.3.4, we did a performance test by varying the duration of noncoding and coding regions, from [100, 400] and [200, 800] respectively, in a step size of 10 nucleotides That is a total combination of 1891, which equals to 31 x 61. The experiment is run on two model microbial genomes, *E. coli* K12 and *B. Subtilis*, both with strong RBS signals.

Taking (Sn+Sp)/2 as the benchmark, Table 3.7 lists top 10 highest accuracy achieved by different combinations noncoding duration and coding duration (ndec and cdec in the table), sorted in descending order of the average value. Note that in these two genomes, the highest accuracy was achieved by different ndec and cdec.

Table 3.7 Best 10 accuracy achieved by varying coding and noncoding duration parameters.

a) *E. coli*

ndec	cdec	# anno	# pred	Sn	Sp	Avg.
260	750	4131	4180	97.02	95.89	96.46
250	780	4131	4178	97.00	95.91	96.46
230	800	4131	4181	97.02	95.86	96.44
250	770	4131	4179	97.00	95.88	96.44
240	800	4131	4179	97.00	95.88	96.44
270	740	4131	4177	96.97	95.91	96.44
260	770	4131	4177	96.97	95.91	96.44
260	760	4131	4177	96.97	95.91	96.44
250	800	4131	4177	96.97	95.91	96.44
250	790	4131	4177	96.97	95.91	96.44

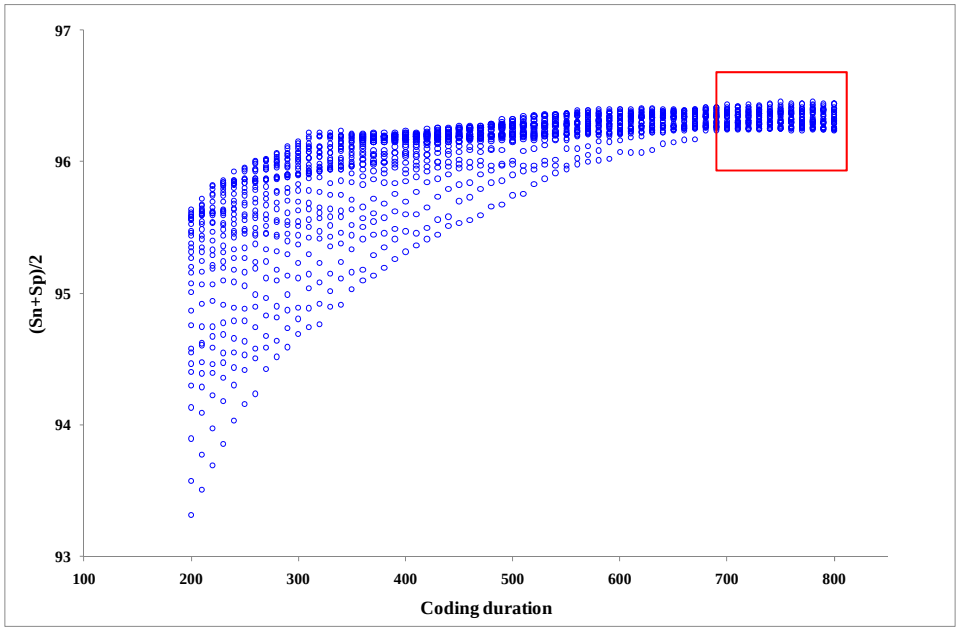
b) *B. subtilis*

ndec	cdec	# anno	# pred	Sn	Sp	Avg.
400	800	4105	4172	97.66	96.09	96.88
390	800	4105	4172	97.66	96.09	96.88
400	790	4105	4172	97.66	96.09	96.88
380	800	4105	4173	97.66	96.07	96.87
340	800	4105	4184	97.78	95.94	96.86
360	790	4105	4182	97.76	95.96	96.86
350	790	4105	4184	97.78	95.94	96.86
370	780	4105	4182	97.76	95.96	96.86
360	780	4105	4184	97.78	95.94	96.86
350	780	4105	4184	97.78	95.94	96.86

There are two parameters available for optimization. With all the possible combinations readily available, it is necessary to fix one at a time to see the overall effect, illustrated in Figure 3.16 a) and b), by fixing noncoding and coding durations, respectively.

As shown by the red box, a long decay (>700nt) duration of coding model gives small variation in the average accuracy measure. The default noncoding decay of 150 and coding decay of 300 was derived from *E. coli* genomes. This setting is probably good for native model from GeneMarkS training. However, the heuristic model is much atypical/ universal. In other words, the model is more relaxed when classifying ORF as gene. It is possible to compensate this by extending the coding decay duration, and then less short ORFs would be predicted, leading to an increased specificity.

a)



b)

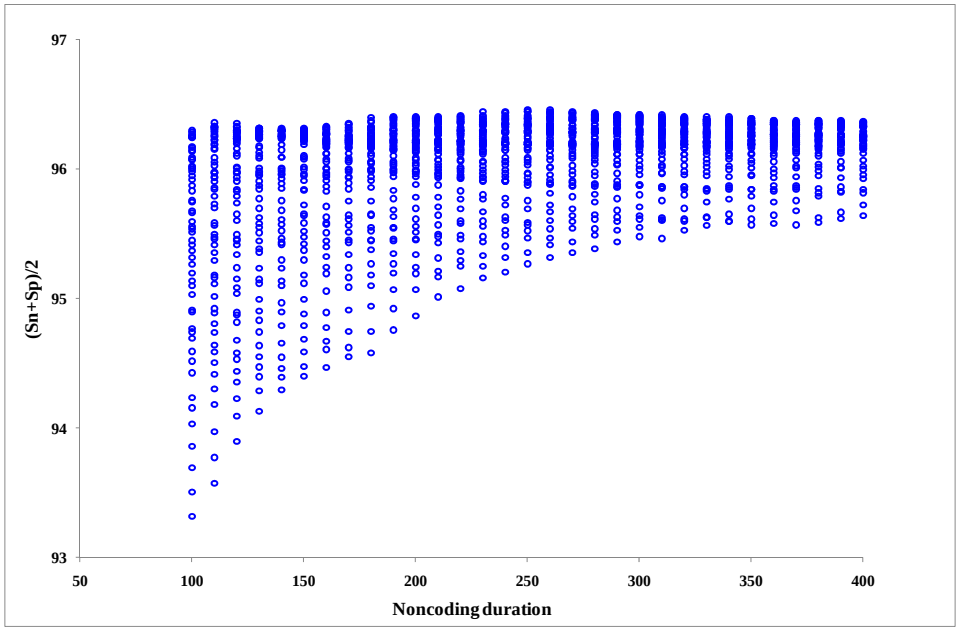


Figure 3.16 Average accuracy by varying duration parameters in *E. coli* K12 genome.

After all, Figure 3.17 shows the joint distribution of sensitivity-specificity. This is similar to what we have done in the metagenomic sequences. The resulting shape forms a curve, and the most upper-right corner data point would be the optimal. Note that the highest sensitivity of 98.69% is paired with a specificity of only 89.62%, an over-prediction of 4549 genes mostly arisen from a low coding duration parameter of cdec 250.

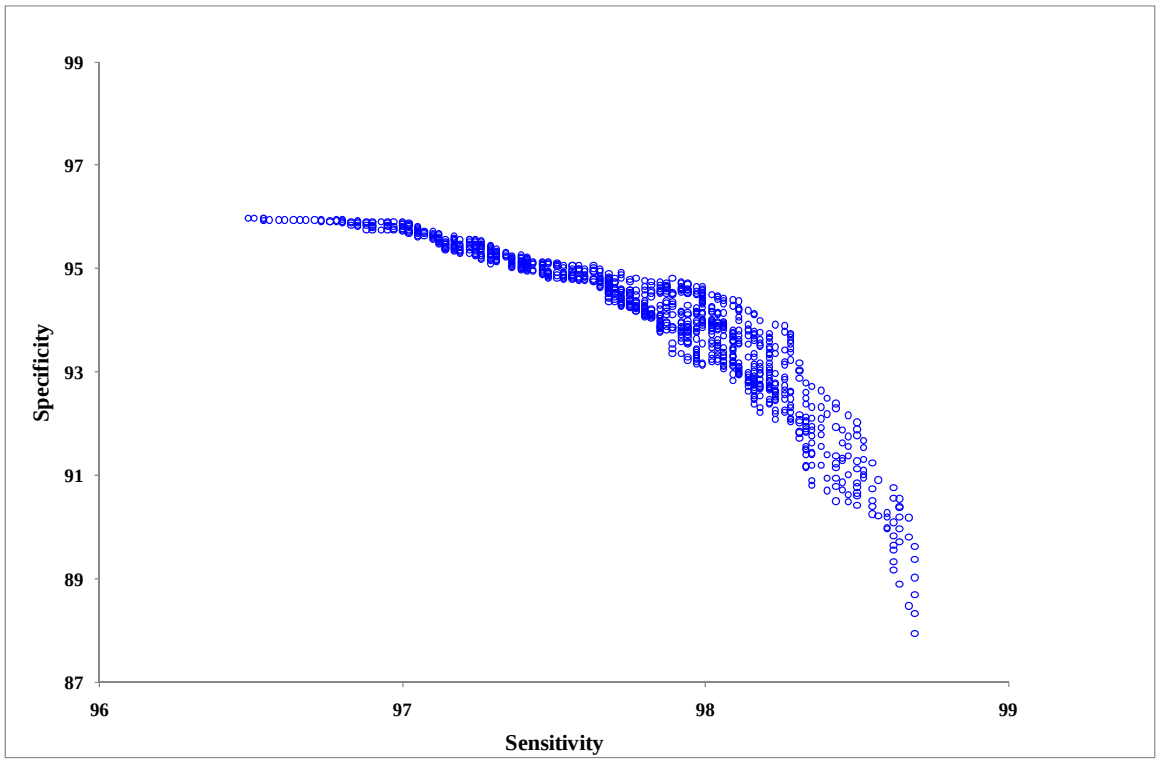


Figure 3.17 Pairs of Sn-Sp achieved by different duration parameters, *E. coli* K12 genome.

Overall, the sensitivity and specificity could range [96.49, 98.69] and [87.94, 95.98], respectively.

3.4.6 GeneMarkS accuracy test

By default, GeneMarkS predicts genes using the combined model, the first iteration heuristic model and the native model from the iterative training. It has been shown that

the addition of the 1999 HAL heuristic model helps to catch atypical codon usage genes (Lukashin and Borodovsky 1998; Besemer, Lomsadze et al. 2001). Other default settings are: RBS mode is ON and the durations of noncoding and coding are 150 and 300, respectively.

We tried to address the following three questions:

1) Compare GeneMarkS or Glimmer 3?
2) What is the effect of substituting 1999 HAL with the new heuristic C-3BA model?
3) What is the effect of using longer coding duration?

We used a test set of 912 bacterial genomes. 50 of these genomes were used as the test set for the metagenomics project. Among which are 34 bacteria and 16 archaea genomes and we call them old genomes. These genomes were sequenced earlier and are thought of as more reliable.

We used the regular measures, the sensitivity and specificity for correct 3' and exact for correct 5' matches. For the sake of conciseness, only the average numbers over the a) 50 old b) 34 old bacteria c) 16 old archaea and d) all 912 genomes were reported.

3.4.6.1 Compare GeneMarkS and Glimmer 3

The authors of Glimmer reported that the version 3 achieves equal or higher sensitivity than version 2, while improves the specificity by reducing false positives (Delcher, Bratke et al. 2007). Table 3.8 shows the difference of accuracy of the GeneMarkS minus that of the Glimmer 3. GeneMarkS is about 0.5 to 1% better than Glimmer 3 in terms of sensitivity and specificity, but much better (10%) in the 5' start calling, under the

circumstances that the Glimmer version 3.0 has implementation of the RBS refinement similar to the way GeneMarkS applied.

Table 3.8 Compare gene prediction accuracy by GeneMarkS 4.6 and Glimmer 3.0.

GeneMarkS - Glimmer 3	Sn	Sp	Sum	Exact
Selected 50 genomes	0.48	0.73	1.21	7.88
Bacteria subset (34)	0.51	1.02	1.53	9.26
Archaea subset (16)	0.48	0.64	1.13	13.43
All genomes (912)	0.10	1.23	1.32	8.64

3.4.6.2 Compare the sub-model, 1999 HAL and 2007 C-3BA

1999 HAL model was a pretty concrete milestone project. Using only 17 genomes, it performs still reasonably well, in predicting atypical genes as well as typical ones. While the 2007 C-3BA codon model was fitted on significantly more genomes (319 bacteria and 38 archaea), the goal of this new project is to find a native model for universal gene prediction.

Table 3.9 Difference in accuracy by the incorporating C-3BA model rather than 1999 HAL.

New heuristic - default	Sn	Sp	Sum	Exact
Selected 50 genomes	-0.48	0.54	0.06	-0.59
Bacteria subset (34)	-0.64	0.61	-0.03	-0.75
Archaea subset (16)	-0.24	0.51	0.27	-0.63
All genomes (912)	-0.62	0.52	-0.10	-0.77

Table 3.9 shows the accuracy difference by substituting the 1999 HAL with the C-3BA model in the combined model of the GeneMarkS output. Sensitivity drops down and this confirms that 1999 HAL is quite capable of finding atypical genes, i.e., those genes with relative low coding potential. Note that the combination of native and 2007 C-3BA model predicts less number of genes. For example, *E. coli* K12 has 4131 annotated genes;

comparing to the 4299 by 2007 C-3BA combined GMS *versus* 4389 by 1999 HAL combined GMS. Roughly, the result shows the cancelling effect of dropping sensitivity and increased specificity.

3.4.6.3 Compare the extended coding duration (800) and the default one (300).

Table 3.10 shows the difference of GeneMarkS with longer coding duration (800) minus GeneMarkS with default coding duration (300). Adding up the sensitivity and specificity, it is a gain of 2.0%. The other necessary clarification is about the "exact match" benchmark. We calculate this in two steps: 1) first compare the 3' stop codons between the annotation and the prediction; 2) Only for those matches, we compare the 5' start codon positions. Now consider the old 50 genomes, the sensitivity drops 0.97% and the exact drops 0.66%. The 3'-end of those 0.97% ORFs were not predicted at all. In order to calculate the true exact match, we need to add back these 0.97% onto the -0.66%, and it turns out a 0.31% gain in this case. This finding is consistent among all other subsets.

Table 3.10 Effect of extending coding duration parameter.

Longer duration - default	Sn	Sp	Sum	Exact
Selected 50 genomes	-0.97	3.07	2.10	-0.66
Bacteria subset (34)	-0.93	2.96	2.02	-0.63
Archaea subset (16)	-1.20	3.23	2.03	-0.73
All genomes (912)	-1.34	3.41	2.08	-0.94

3.4.7 Post-processing with TriTISA

Our collaborator, a team led by Dr. Huaiqiu Zhu, has developed several translation initiation site correction programs (Zhu, Hu et al. 2007; Hu, Zheng et al. 2008; Hu, Guo et al. 2009). Their idea stemmed from the RBS module embedded in GeneMarkS. They

reported that the programs are robust and do not depend on the input, which is the output from other gene finding programs (Hu, Zheng et al. 2009). They score the prestart region using two positional weight matrices of Shine-Dalgarno sequence and TATA-box. A Bayesian method (Hu, Zheng et al. 2008) was employed to predict the potential TIS by taking the highest score one from the candidates. Later in 2008, TriTISA improved by applying a higher order Markov model to train the PWM parameters.

In our test, TriTISA was used as a post-processing on the predictions made by GeneMarkS. Table 3.11 lists its performance on three well known data sets. EcoGene (Rudd 2000) contains 858 5' experiment verified genes. Improvement was observed in both of the *E. coli* set, but a marginal drop was seen on the *B. subtilis* genome.

Table 3.11 TriTISA performance on three data sets.
Improvement is shown in bold font.

Data set	# of annotated genes	3' found	5' correct	TriTISA
EcoGene	858	856 (99.77%)	805 (93.82%)	**817 (95.22%)**
E. coli genome	4131	4059 (98.26%)	3621 (87.65%)	**3726 (90.20%)**
B. subtilis genome	4105	4057 (98.83%)	3540 (86.24%)	3529 (85.97%)

We further tested on all genomes, trying to assess TriTISA's performance in large scale. To our surprise, GeneMarkS is already comparable with TriTISA, even has a 3.03% advantage in those 16 old archaeal genomes (Table 3.12).

Table 3.12 Result of TriTISA, based on GeneMarkS predictions

GeneMarkS	Sn	Sp	Exact	TriTISA improvement
Selected 50 genomes	97.11	92.91	76.32	-0.19
Bacteria subset (34)	96.86	93.24	77.51	0.48
Archaea subset (16)	98.84	93.03	81.77	-3.03
All genomes (912)	96.62	91.45	78.63	-0.41

3.5 Other aspects that could help

There are three items yet tried or implemented in current GeneMarkS version.

a) Pseudogenes finding

It is similar to the problem of tRNA genes, in the sense that pseudogenes directly lead to false positive CDS calling. Meanwhile, it is relatively simple, once we have knowledge about their locations from extrinsic information (Lam, Khurana et al. 2009). Under usual circumstances, pseudogenes in a particular genome may not be many enough to bias the HMM model parameter training. However, special care must be taken for particular genomes, such as *Mycobacterium leprae*, which has 1116 genes within a 3.27MB genome (Cole, Eiglmeier et al. 2001).

b) G+C content heterogeneity

Variation and heterogeneity of DNA base composition has long been a research subject (Sueoka 1962). Despite the fact that GC variation mostly exists in eukaryotic genomes (Nekrutenko and Li 2000), *Xylella fastidiosa* was reported to be one of several prokaryotic genomes which are markedly heterogeneous in DNA composition (Bernaola-Galvan, Oliver et al. 2004). We have tried to manually evaluate the GC skew and then applied the corresponding heuristic model corresponding to local GC-content with success. The next step could be define more hidden states inside the Markov chain, so that the model could be selected depending on the GC content of the regions being analyzed.

c) Operon structure and frame shift detection

Operon structure could be further used. The spacer length distributions of the first gene and the rest genes' are so different that it could be modeled to help refine the gene start prediction.

The other issue is about the frame shift. GeneTack (Antonov and Borodovsky 2010) has proved to be effective. It will need careful investigation to use GeneTack as i) a pre/post processing module for GeneMarkS or ii) incorporate into the HMM structure, which has the possibility to lead to false predictions.

* This chapter was the preliminary result for the following publication in preparation (Zhu, Lomsadze et al. 2010):

Zhu W., Lomsadze A. and Borodovsky M.

GeneMarkS Plus: Improving gene annotation in complete prokaryotic genomes.

In Preparation.

CHAPTER 4 Codon usage and expression level analysis in *Bacillus anthracis* genome

Abstract

Earlier studies showed that the codon usage bias and tRNA copy numbers contribute to the protein translation optimization. The new SoLiD sequencing data enabled us to use the whole transcriptome shotgun sequencing of a bacterial pathogen *Bacillus anthracis* to assess correlation of gene expression level with codon adaptation index, RBS scores, as well as with a new measure of gene translational efficiency, average translation speed. Transcriptome mapping may also improve existing gene annotation. Upon assessment of accuracy of current annotation of protein-coding genes in the *B. anthracis* genome we have shown that the transcriptome data indicate existence of more than a hundred genes missing in the annotation though predicted by an *ab initio* gene finder. Also, we compared computational predictions of operon topologies with the transcript borders inferred from RNA-Seq reads. The results show that the new ATS index, the average translation speed of a gene, as well as CAI correlate with gene expression level. Moreover, contrary to what was thought before, we found a correlation of the score of an RBS site with gene expression level of the downstream gene for genes that appear to be the first genes in operons.

4.1 Introduction

The codon adaptation index (CAI) is a widely used numerical index and suggests the relationship between gene expression level and codon usage bias. Rocha *et. al* reported that a small subset of optimal codons are dominant in highly expressed genes across different fast-growing bacteria (Rocha 2004). Interestingly, the synonymous codon most frequently used in its group as defined for the whole gene complement is not always the same as the synonymous codons most frequently used in its group in a subset of genes with high expression. We have introduced a new measure of efficiency of translation, the average translation speed of a gene, the ATS index. This chapter addresses the following questions. 1) What is the correlation of CAI and ATS values with gene expression level? 2) What is the correlation between two neighboring same-strand genes?

4.2 Materials

The complete *Bacillus anthracis* Ames Ancestor genome sequence and its RefSeq annotation (NC_007530) were downloaded from NCBI. The genome is 5,227,419 nucleotides in length and has a GC content of 35.4%. We used a set of *B. anthracis* candidate genes predicted by GeneMarkS (Besemer, Lomsadze et al. 2001). This genomic gene pool consists of a total of 5,661 genes. In order to calculate the codon frequencies, we removed the start codons (ATG, GTG and TTG) and stop codons (TAG, TGA and TAA) of every gene. Genomic codon usage frequencies were derived by concatenating all the predicted genes.

The transcriptome RNA-Seq data from four growth stresses conditions was analyzed. These four conditions are (i) cold shock; (ii) osmotic shock as imposed by 0.75M sodium chloride (NaCl); and (iii) 6% ethanol shock. For each condition, the expression of gene was estimated in the following way. The expression level of each *B. anthracis* gene can be estimated by counting the number of SOLiD reads mapped to a gene normalized by the gene's length. The number of reads mapped to a given gene can be accurately measured given only the read counts within the gene, assuming a read length of 35 nt. For the purposes of correlation, this strategy provides a good estimate of gene expression level. The log-base 2 of the maximum expression level for a given gene across all conditions was used in the subsequent correlation analysis.

4.3 Methods

All the annotated genes were concatenated to count the genome-wide codon and to derive the codon usage frequency.

Revisiting the codon adaptation index (CAI)

The relative adaptiveness of a codon is defined as its frequency relative to the most often used synonymous codon, which is computed from a set of highly expressed genes ***G***.

$w_{aa,i}(G) = \frac{f_{aa,i}}{f_{aa,max(G)}}$, where $f_{aa,i}$ is the frequency of codon i which encodes amino acid aa, and $f_{aa,max}$ is the frequency of the codon most often used in a set of highly expressed genes ***G***. The CAI of a particular gene g is simply the geometric average of the relative adaptiveness in a gene sequence.

$CAI_g = \prod_{i=1}^{N} w_i^{1/N}$, where w_i is the relative adaptiveness of the i-th codon in a gene with ***N*** codons. (Jansen, Bussemaker et al. 2003). Both w_i and CAI values ranges from 0 to 1,

with 0 indicating that a codon is not present at all while 1 indicating the codon occurs most often for a given amino acid.

To define gene set ***G***, the authors of the original publication used a set of only 27 highly expressed *E. coli* genes and calculated the codon composition of such set. (Sharp and Li 1987)

Measures of translational efficiency

As a predictor of translational efficiency for an mRNA we introduced an average translation speed (ATS) defined as follows. Let frequencies of 61 codons in a reference gene set be s_i, i = 1,2, …, 61. Since evolutionary adaptation of the codon and anti-codon (tRNA) populations is supposed to eliminate disproportions at a time of fast growth, we assume that the frequencies of tRNA in a cell are proportional to s_i values. Before a cognate tRNA is admitted to the A site at a ribosome, a number of candidate tRNA are tried and rejected. We assume a Poisson process for interactions between a cognate tRNAs and the ribosome A site; thus, the average time needed for recruiting a cognate tRNA is proportional to $1/s_i$. For a gene with N codons and k_i codons of each kind the average time of mRNA translation is $T = \sum k_i / s_i$. Then, for a given gene the average time of a codon translation is $t = T/N = \sum (k_i /N)/s_i$. Finally, with k_i /N being a frequency of a codon i in the gene, designated as f_i, we have $t = \sum f_i /s_i$ and the average speed of translation of the gene is $V = (\sum f_i /s_i)^{-1}$. More accurate computation of the average speed of codon translation requires data on concentration of each mRNA, knowledge that has not been available until recently. In this study we use the RNA-Seq derived information on gene expression levels observed in *B. anthracis* (see below) to make correction in the s_i values. Instead of s_i defined as $\sum \mu^j_i / \sum\sum \mu^j_i$ for each codon type i among the genes in the reference set, with μ^j_i being a count of codons of type i in gene j, we used the formula,

$S_i = \sum w_j \mu^j_i / \sum\sum w_j \mu^j_i$ where w_j is the expression level of gene j. Now the formula for V can be modified and we defined the value of ATS index, the average translation speed of a gene, ATS = $(\sum f_i / S_i)^{-1}$. For comparison, we also used the classic CAI measure defined by by Sharp and Li (Sharp and Li 1987).

4.4 Results

4.4.1 GeneMarkS prediction on B. anthracis

As a sanity check, we compared the RefSeq annotation and the GeneMarkS prediction. Table 4.1 shows the differences. GeneMarkS detected strong RBS signal AGGAGG in this low GC content (35.4%) genome, as well as localization signal centered at 8 nucleotides upstream of the start codons, as illustrated in Figure 4.1.

Table 4.1 Comparing RefSeq annotation (NC_007530) and prediction by GeneMarkS combined model, with RBS option turned on.

# annotation	# prediction	Sn %	Sp %	Exact %
5308	5661	96.91	90.87	85.14

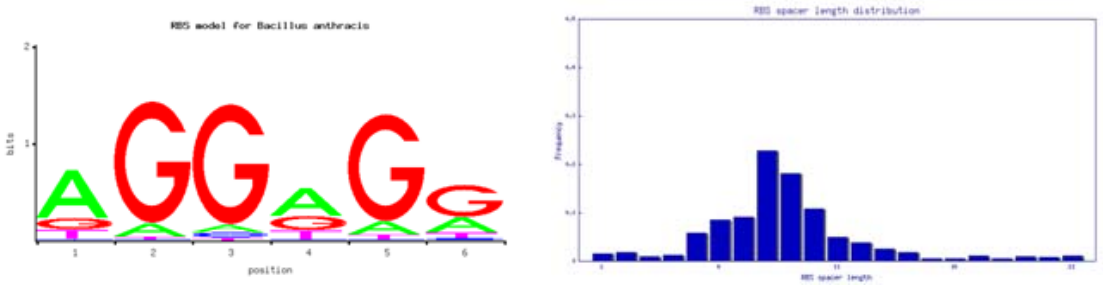


Figure 4.1 RBS site of *B. anthracis*.

4.4.2 tRNA gene type/abundance and the protein-coding genes expression level

The RefSeq annotation (Accession#: NC_005730) does not give the anti-codon information for the annotated tRNA genes. As an alternative, we used the state-of-the-art tRNA gene finding program, tRNASCAN-SE (Lowe and Eddy 1997) for detection. By doing the reverse and complement of the anti-codon found, we constructed the Table 4.2, listing the 95 tRNA genes found. The most frequent used codon in each synonymous group is marked with bold font. Quite many codons are listed as zero. This finding confirmed the Wobble hypothesis proposed by Francis Crick, back in 1966 (Crick 1966). The hypothesis stated that, the 5' base on the anticodon, which binds to the 3' base on the mRNA, was not so confined as the other two bases. Thus, some tRNA species could pair with more than one codon. Moreover, Kato *et. al* showed that the mismatching at the wobble pair position does reduce translation efficiency (Kato, Nishikawa et al. 1990).

Gene expression data delivered by mapped RNA-Seq reads allows for ranking genes by expression levels. For the sake of comparison, we have increased the size of the reference set to 100 genes, with 48 of these genes coding for ribosomal proteins (Supplementary Table 11). A comparison of codon frequencies in the whole complement of genes and in the 100 most highly expressed genes (under Control condition) shows that seven synonymous groups have different optimal codons. The list of optimal codons is interesting to compare with the list of tRNA genes (Table 4.2). In 6 out 18 cases the optimal codon in 100 highly expressed genes does not match the exact tRNA species present in the *B. anthracis* cell; the optimal codon in the whole gene complement does not match the exact tRNA species in 9 out of 18 cases.

Table 4.2 *Bacillus anthracis* codon frequencies in the whole set of genes and several gene subsets and the copy numbers of tRNA genes.
The codon frequencies were calculated from three sets of coding sequences, namely global genome, 100 the most highly expressed genes as observed from the RNA-Seq data and the 37 homologs to the proteins which used by Sharp et.al (1987). The codon frequencies are normalized to 1000. The "Weighted 100 genes" column shows the frequencies of codons adjusted by expression levels of the 100 genes weighted as determined from the RNA-Seq data, in order to approximate the whole population of codons and calculate the ATS value. 95 tRNA genes were assigned to codons by using tRNAscan-SE and are shown in "tRNA genes" column. Numbers in bold font indicate the maximum frequencies/counts in a synonymous group of codons.

Amino acid	Codon	Genomic Freq	100 genes	Weighted 100 genes	37 genes (CAI)	tRNA genes	Amino acid	Codon	Genomic Freq	100 genes	Weighted 100 genes	37 genes (CAI)	tRNA genes	Amino acid	Codon	Genomic Freq	100 genes	Weighted 100 genes	37 genes (CAI)	tRNA genes	Amino acid	Codon	Genomic Freq	100 genes	Weighted 100 genes	37 genes (CAI)	tRNA genes
Phe	**TTT**	**32.8**	10.2	9.3	5.1	0	**Ser**	**TCT**	**15.5**	**23.5**	**24.1**	**26.1**	0	**Tyr**	**TAT**	**28.0**	10.1	9.2	6.2	0	**Cys**	**TGT**	**6.3**	**3.4**	**2.8**	**2.9**	0
	TTC	14.4	**23.6**	**22.7**	**21.6**	**4**		**TCC**	3.2	1.1	1.1	0.3	1		**TAC**	9.3	**15.5**	**14.8**	**14.6**	**2**		**TGC**	2.1	1.4	1.3	2.0	**1**
Leu	**TTA**	**49.9**	**42.1**	**40.1**	**38.2**	**2**		**TCA**	14.7	8.9	8.6	7.3	**4**	**Stop**	**TAA**	0.0	0.0	0.0	0.8	0	**Stop**	**TGA**	0.0	0.0	0.0	0.3	0
	TTG	9.2	3.2	2.9	1.1	1		**TCG**	4.6	1.3	1.2	0.9	0	**Stop**	**TAG**	0.0	0.0	0.0	0.6	0	**Trp**	**TGG**	**10.4**	**6.2**	**5.2**	**4.0**	**2**
Leu	**CTT**	18.2	22.7	22.8	26.2	0	**Pro**	**CCT**	9.1	9.9	10.6	11.5	0	**His**	**CAT**	**16.4**	**8.6**	7.7	7.8	0	**Arg**	**CGT**	**14.1**	**36.8**	**40.7**	**51.2**	**3**
	CTC	4.2	0.8	0.8	0.2	1		**CCC**	1.1	0.2	0.3	0.2	0		**CAC**	4.8	8.1	**8.3**	**9.6**	**2**		**CGC**	4.8	9.5	10.0	12.4	0
	CTA	10.6	8.5	7.6	6.4	2		**CCA**	**16.5**	**23.2**	**22.1**	**20.2**	**3**	**Gln**	**CAA**	**30.3**	**30.7**	**30.0**	**31.3**	**4**		**CGA**	5.4	1.2	0.9	1.6	0
	CTG	3.3	1.9	1.7	1.2	0		**CCG**	7.5	2.9	2.5	1.6	0		**CAG**	6.7	3.7	4.0	2.5	0		**CGG**	1.3	0.1	0.0	0.3	1
	ATT	**50.8**	30.5	27.6	24.5	0	**Thr**	**ACT**	12.4	26.0	**29.9**	**28.4**	0	**Asn**	**AAT**	**32.9**	14.8	14.2	11.0	0	**Ser**	**AGT**	14.5	5.9	4.8	4.5	0
Ile	**ATC**	13.4	**32.7**	**35.3**	**36.3**	**4**		**ACC**	2.5	0.7	0.6	0.3	1		**AAC**	13.2	**28.6**	**28.7**	**27.2**	**5**		**AGC**	5.8	6.5	6.2	4.8	2
	ATA	16.9	1.8	1.3	1.4	0		**ACA**	**27.9**	**28.4**	26.9	25.1	**4**	**Lys**	**AAA**	**56.1**	**62.8**	**66.7**	**72.0**	**5**	**Arg**	**AGA**	9.4	3.1	3.3	4.2	1
Met	**ATG**	**25.0**	**24.7**	**22.9**	**21.4**	**8**		**ACG**	13.5	5.4	5.3	3.9	0		**AAG**	18.1	14.1	15.4	17.7	0		**AGG**	2.4	0.2	0.2	0.6	0
	GTT	25.9	38.5	41.1	**42.8**	0	**Ala**	**GCT**	21.0	**41.9**	**46.2**	**54.6**	0	**Asp**	**GAT**	**37.8**	**29.8**	**28.3**	**27.2**	0		**GGT**	**24.9**	**49.1**	**51.3**	**47.5**	0
Val	**GTC**	5.7	2.7	2.4	1.6	1		**GCC**	4.0	1.4	1.3	0.8	0		**GAC**	8.8	18.9	19.7	21.9	**6**	**Gly**	**GGC**	8.4	12.2	11.6	11.3	**4**
	GTA	**31.1**	**43.3**	**43.6**	39.6	**5**		**GCA**	**29.7**	38.5	35.1	32.6	**5**	**Glu**	**GAA**	**57.2**	**63.2**	**63.8**	**66.3**	**7**		**GGA**	24.4	18.6	17.8	16.1	**4**
	GTG	10.5	5.8	5.9	6.5	0		**GCG**	13.1	9.9	8.8	8.7	0		**GAG**	18.5	17.9	18.4	21.6	0		**GGG**	9.5	2.7	2.2	1.7	0

4.4.3 The effect of selecting reference set

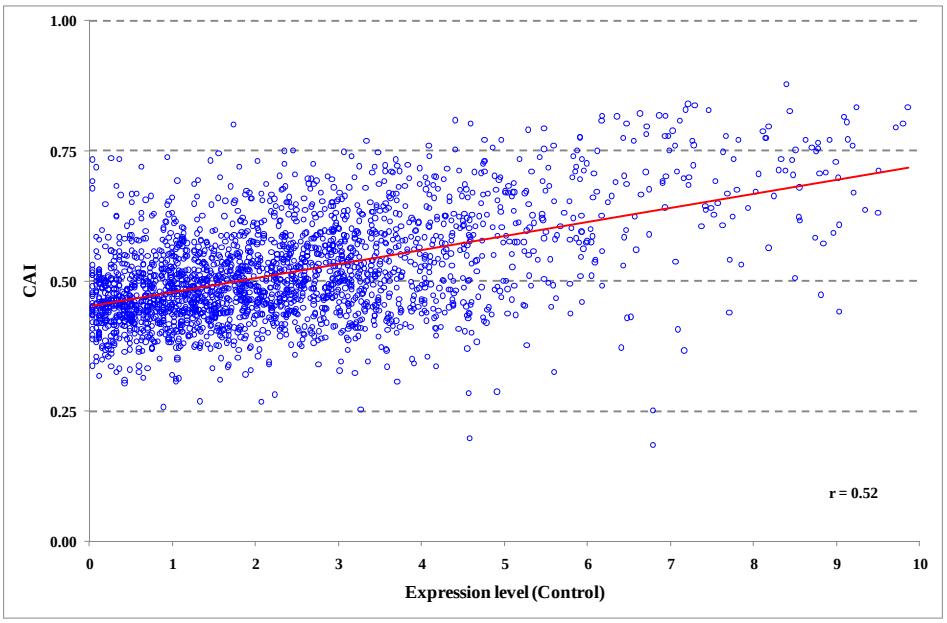
The CAI values depend on the codon composition of the putative highly expressed gene set ***G***. It is a question how to select this set in the gene pool of *B. anthracis*, using either the gene pool from the whole genome or the subset, the ribosomal genes as it was done by Sharp *et. al* in 1987.

The values of codon adaptation index and average translation speed for a given gene depend on the model parameters derived from codon frequencies in a selected reference set of genes. In the original paper Sharp and Li (Sharp and Li 1987) used 27 *Escherichia coli* genes with experimentally demonstrated high expression. Obviously, orthologs of these genes in *B. anthracis* could make a reference set for computing CAI values for *B. anthracis* genes. However, several genes in the 27 strong set of *E. coli* genes do not have orthologs in *B. anthracis.* Therefore, we have added several ribosomal protein genes with the same total length, 1555 codons, to make up for the missing genes (a total of 37 genes listed in Supplementary Table 10. Interestingly, codons with highest frequencies (optimal codons) in the groups of synonymous codons, are not the same in the reference set of highly expressed genes and in the whole complement of *B. anthracis* genomic genes (Table 4.2).

We determined the values of codon adaptation index, CAI, for each gene using either 37 or 100 highly expressed genes as a reference set. Similarly, we plotted CAI values, as a function of gene expression level. (Figure 4.2). These figures show almost identical behavior of CAI with respect to a choice of the reference set. We compared the values of CAI and ATS for sets of ribosomal protein genes and genes of transcription factors for two reference sets: 37 and 100 highly expressed genes (Figure 4.3a and b). One can see

that computation of CAI and weighted ATS based on the smaller set of 37 genes provides better separation of the two groups of genes with high and low expression levels.

(a)



(b)

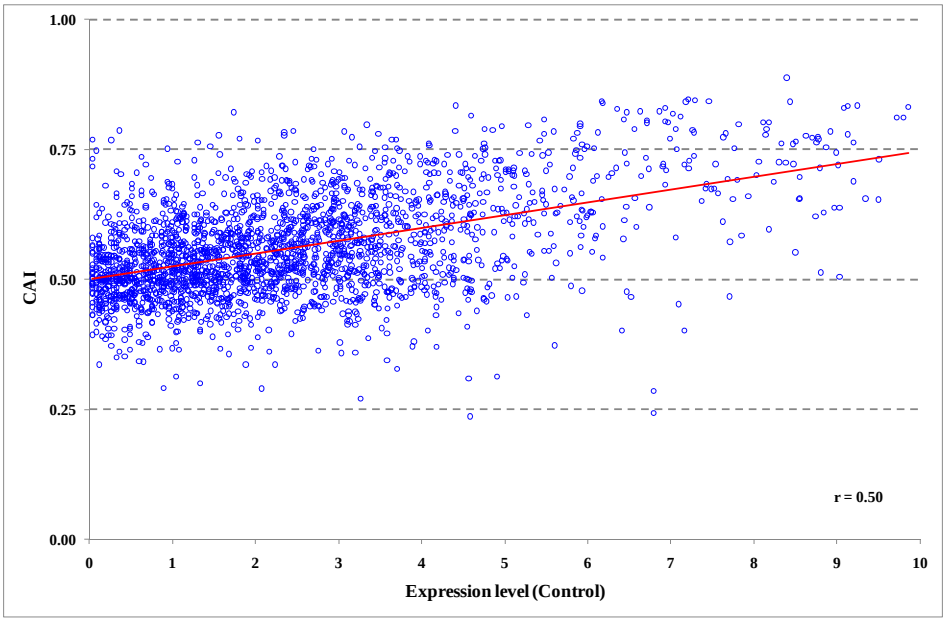
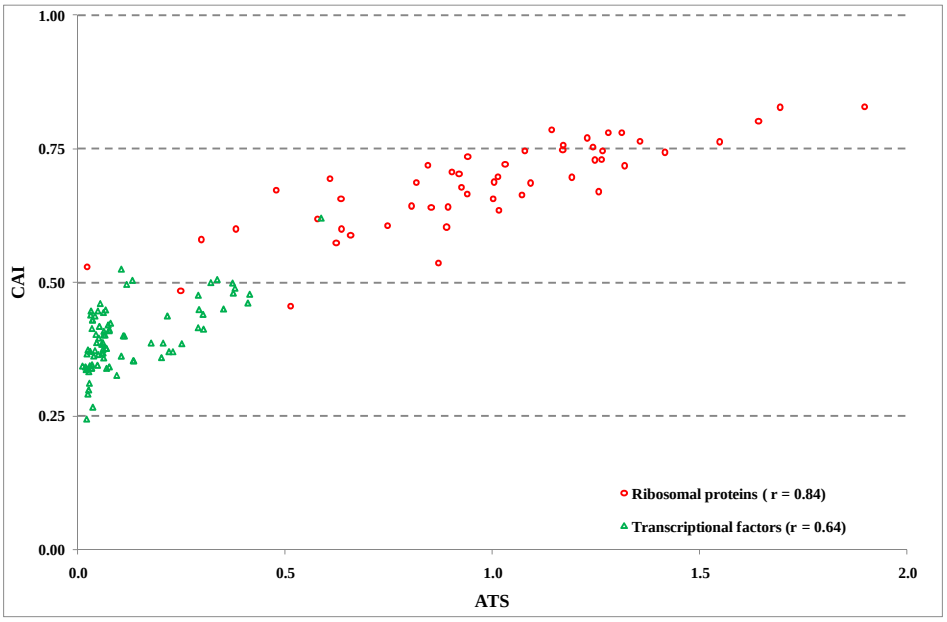


Figure 4.2 Joint distribution of gene expression levels and CAI values.
CAI was calculated using a) 37 proteins homologous to highly expressed *E. coli* proteins (including those selected by Sharp 1987); b) 100 most highly expressed genes inferred from the RNA-Seq data. There is no obvious advantage in using a larger set of genes with high expression.

(a)



(b)

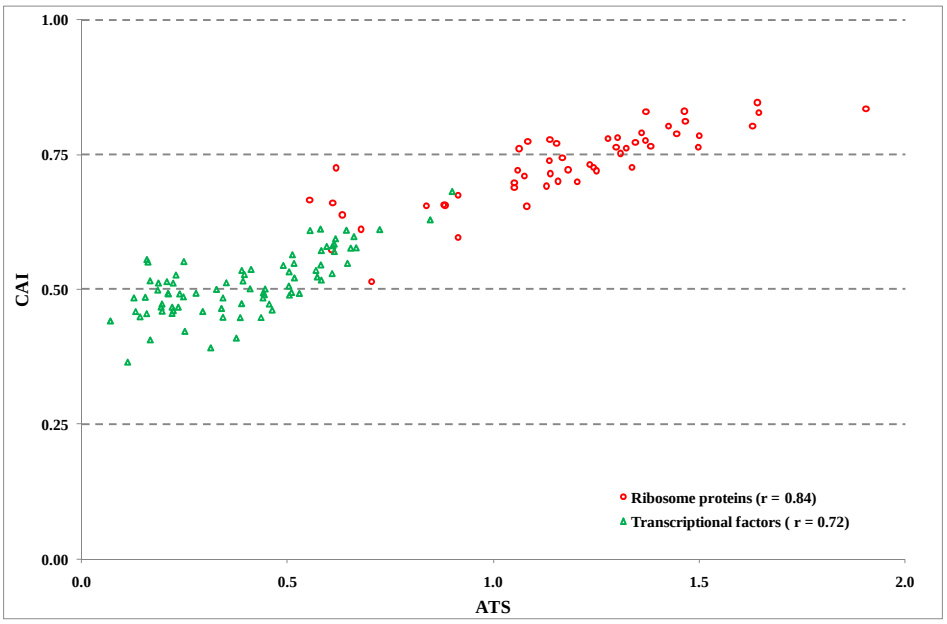


Figure 4.3 Joint distribution of ATS (weighted) and CAI values of *B. anthracis* ribosomal proteins and transcription factors.
ATS and CAI were calculated using a) 37 proteins homologous to highly expressed *E. coli* proteins (including those selected by Sharp 1986); b) 100 most highly expressed genes inferred from the RNA-Seq data.

4.4.4 Correlation of ATS and gene expression level

The RNA-Seq experiment identifies 35 nt fragments of expressed genes. Based on the sequence, we mapped these fragments back onto genomic sequences. In this way, we can count the times of any gene expressed under the experiment condition. By taking the logarithm of these counts with base of 2, we can further select those genes which have a positive value as expressed gene, a total of 2,375 genes out of a total 5661 in *B. anthracis* genome. Figure 4.4 shows the expression level and weighted ATS value of these expressed genes. The correlation coefficient was found to be 0.525.

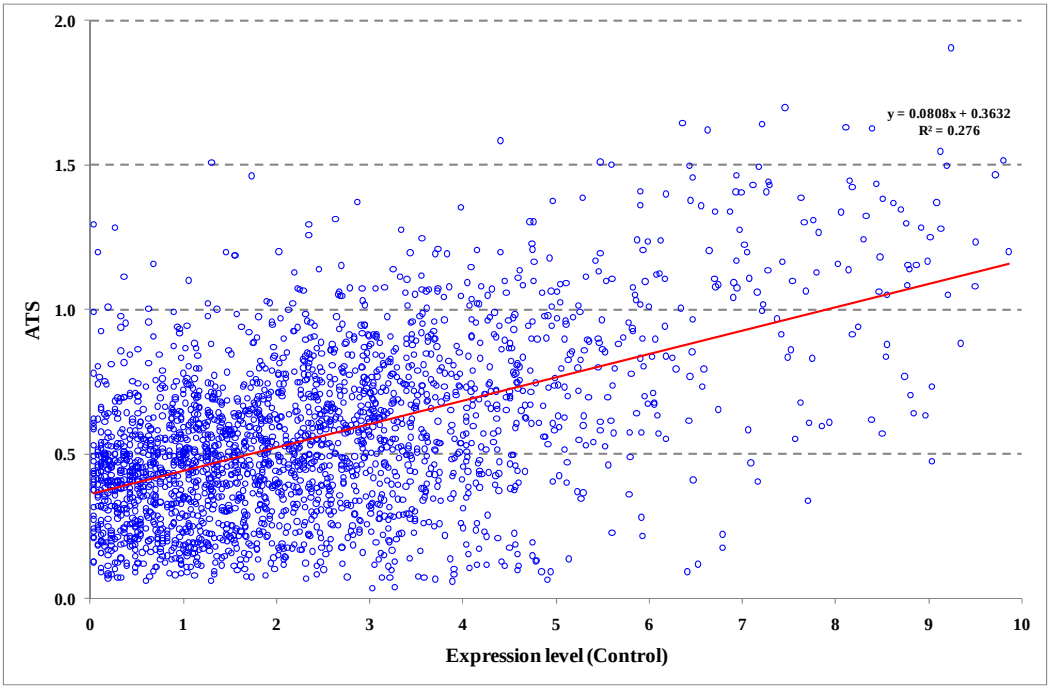


Figure 4.4 Correlation of ATS and gene expression level

4.4.5 RBS score and gene expression level

It is commonly believed that the translational efficiency of prokaryotic mRNAs is intrinsically determined by both primary and secondary structures of their translational initiation regions. We applied GeneMarkS (see Methods section 3.3.3) to score the ribosomal binding sites in prestart region. In order to explore the effect of RBS on the

gene expression level, we tend to select those first genes of operons. We selected genes preceded by non-coding regions longer than 100 nt. From this set we further selected a subset with average coverage by RNA-Seq reads larger than 1, a total of 748 genes. In contrast with earlier observation of no correlation between the RBS score and gene expression level (Lithwick and Margalit 2003) we did observe a weak but significant correlation (Figure 4.5) with correlation coefficient 0.158. This result means that there is a trend for genes with higher expression to have stronger RBS sites. This trend could be expected as genes expressed at high level need to be tightly regulated at all levels including the translation level.

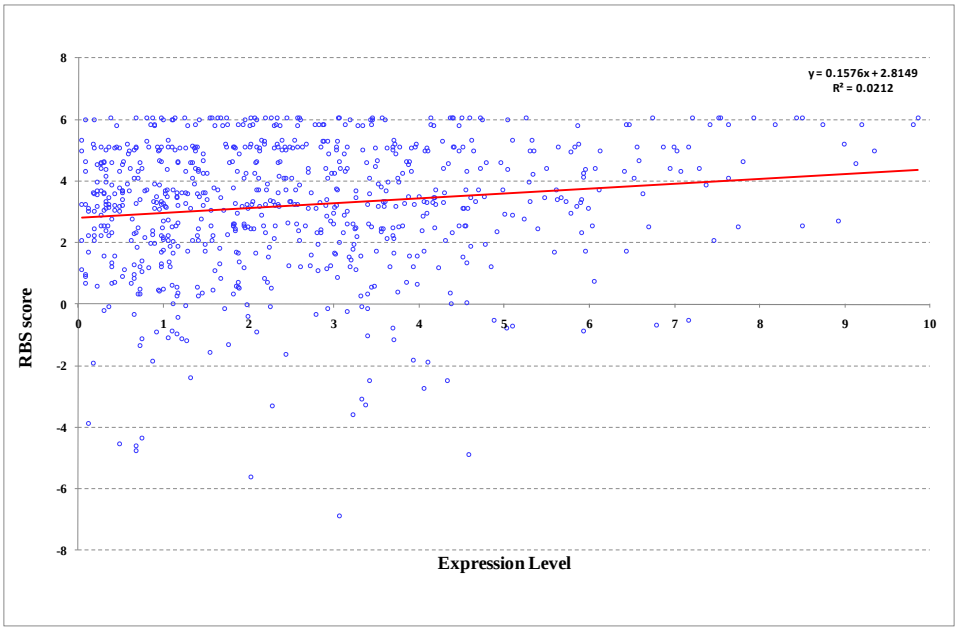


Figure 4.5 RBS score and gene expression level

4.4.6 Correlation of ATS values for pairs of genes with -4, -1 overlaps and separation of more than 100nt

Another focus of our research was to see the correlation of expression level among neighboring gene pairs. Prokaryotic genomes have well-defined operon structure. We

tried to utilize the fact that, the gene expression levels are similar for those genes inside the same operon. For this purpose, we can only consider those neighboring pairs on the same DNA strand. There are 4041 same-strand gene pairs out of 5660 possible ones in *B. anthracis* genome. Figure 4.6 plots these pairs' gene expression level determined by RNA-Seq experiment. The correlation coefficient was 0.67.

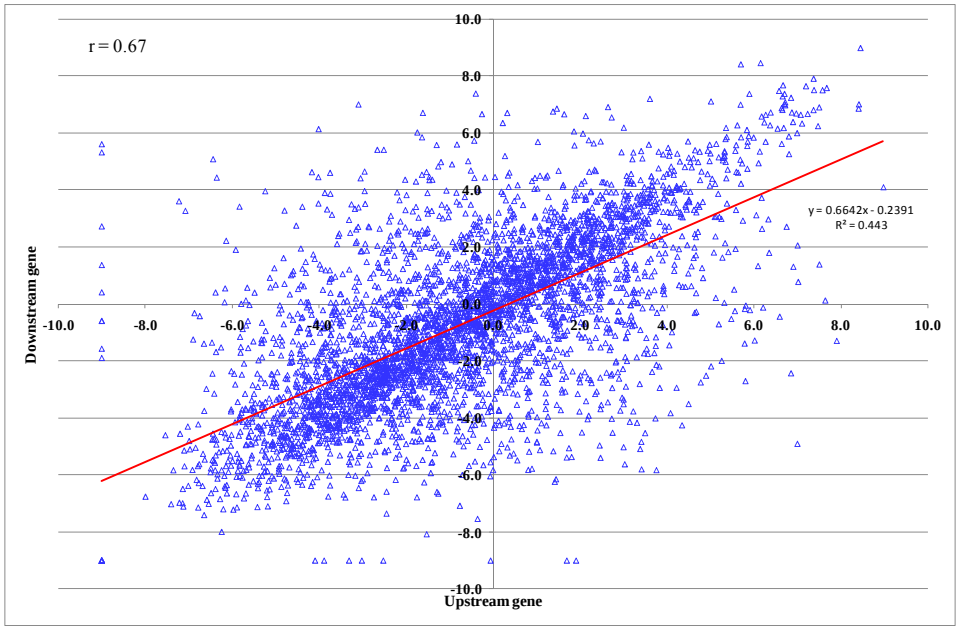


Figure 4.6 Gene expression level of 4041 same-strand gene pairs of *B. anthracis* genome.

Generally speaking, genes inside the same operon are more likely to be in the close proximity (<50nt) to each other (Pertea, Ayanbule et al. 2009). By checking the spacer length between two genes, we categorized these 4041 genes into three subgroups, -4, -1 and more than 100 nucleotides. The total gene numbers of each group are 255, 76 and 1690, respectively. The rest genes were discarded. The minus sign indicates that the two genes overlap each other by several nucleotides, and such pairs are assumed to be in the same operon. On the other hand, those large spacer gene pairs should reside in two

different operons and co-expression is not expected. Table 4.3 lists the correlation coefficients of these three subsets. The overlapping genes have correlation coefficient close to +1, which indicates that they are co-expressed at similar level; while this effect is remarkably reduced for distant gene pairs (correlation coefficient = 0.377).

Table 4.3 Correlation coefficient of gene expression level between two neighboring genes.

Pair Group	Correlation coefficient
-4 nt (255 pairs)	0.942
-1 nt (76 pairs)	0.951
>100nt (1690 pairs)	0.377

* This chapter was part of the following publications (Martin, Zhu et al. 2009; Martin, Zhu et al. 2010):

Martin J., Zhu W., Bergman N. and Borodovsky M. (2009)

Assessment of Gene Annotation Accuracy by Inferring Transcripts from RNA-Seq. BIBM 2009: 54-59

Martin J., Zhu W., Passalacqua K., Bergman N. and Borodovsky M. (2010)

Bacillus anthracis genome organization in light of whole transcriptome sequencing. BMC Bioinformatics 2010, 11(Suppl 3):S10

CHAPTER 5 Gene finding in EST sequences of wheat leaf fungus *Puccinia triticina*

Abstract

We describe an application of GeneMarkS program on the expressed sequence tags (EST) data from fungus pathogen *Puccinia triticina*. The EST sequences were possibly contaminated by the genomic sequences of its plant host. We derived a second order initial Hidden Markov model from all the EST data and made the initial prediction of 11,260 genes on 10,576 EST fragments. The genomic sequence of a close relative species, namely *Puccinia graminis*, was completely sequenced and available. We validated the initial set of predicted proteins by finding similarity to the proteome of *P. graminis*. Based on the resulting validated 2,093 homologous genes, we estimated the parameters of a refined fourth order model and made a second prediction of 7,594 genes. For those EST fragments with one or more genes predicted, we performed a sanity check and found 905 possible frameshifts by conceptually translating the neighboring ORFs. On the other hand, for the rest EST fragments with no genes predicted by the refined model, we applied the heuristic model and made 5,156 gene predictions on a set of 5350 EST fragments. Most of these genes turned out short less than 150 nucleotides, and they were likely the sequence contaminations from the host genome.

5.1 Background information

This was a collaboration project (June. 2009) with Dr. Guus Bakkeren of Departement of Botany, the University of British Columbia.

They had a dataset of sequenced Expressed Sequence Tags (EST) of a rust fungal pathogen species, *Puccinia triticina*. It was possible that these sequences were contaminated by the genomic sequences of its plant host.

In order to identify genes (and protein products), we applied GeneMarkS gene finding program on the EST dataset, refined the predicted protein by BLASTp similarity search against NCBI non-redundant database as well as a close relative species, *Puccinia graminis*, and tried to detect the frame shift by our in-house program GeneTack (Antonov and Borodovsky 2010).

5.2 Materials

There were a total of 13,328 EST sequences, and 2818 of them were shorter than 300 nucleotides, as illustrated in Figure 5.1, the length distribution. On the other hand, the GC content of these sequences were distributed from 30% to 70%, as shown in Figure 5.2, stacked bar graph of short (≤300nt) and relative long (>300nt) fragments.

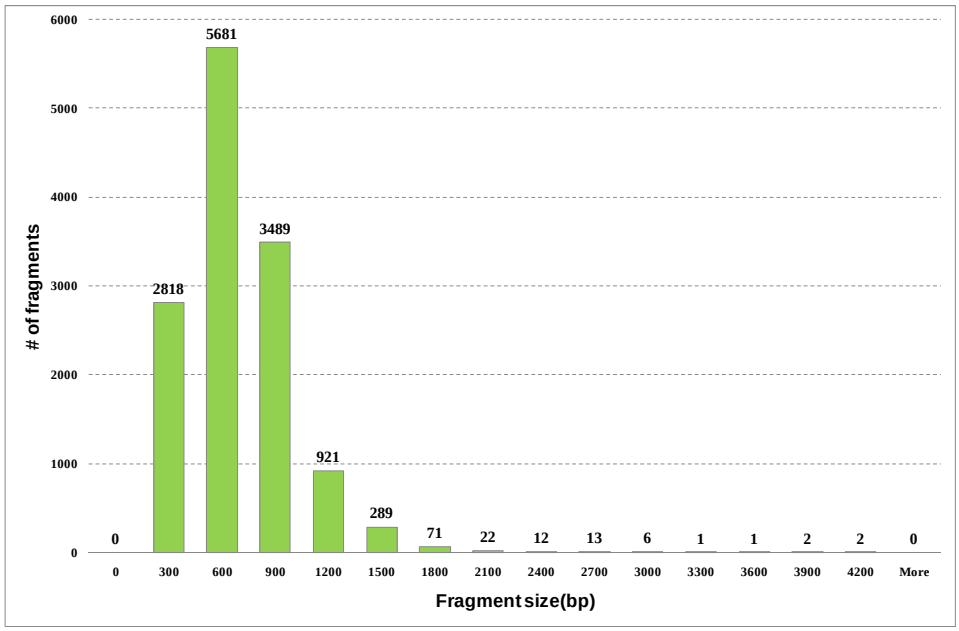


Figure 5.1 EST sequence length distribution

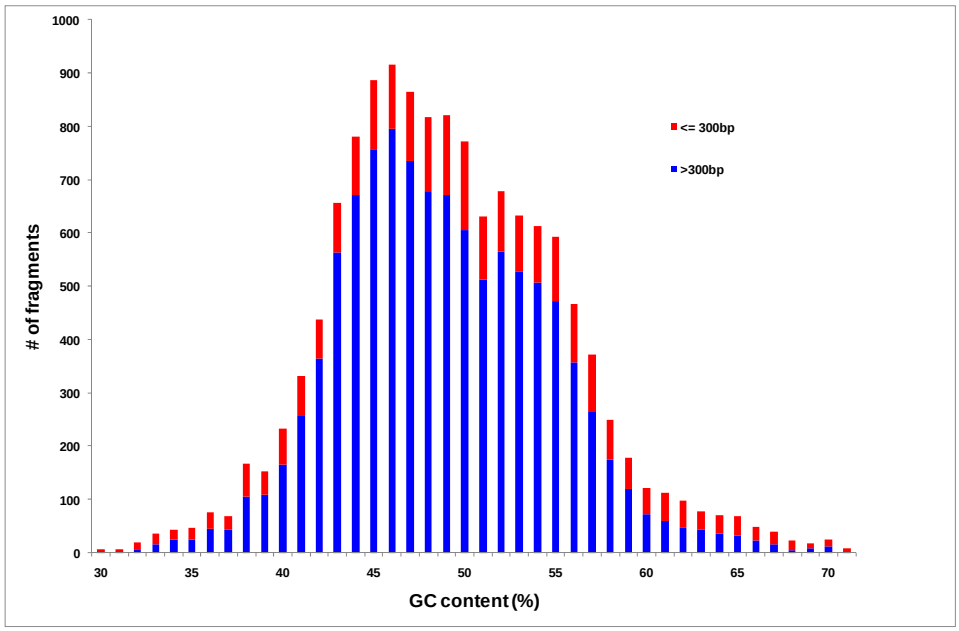


Figure 5.2 GC content distribution of the input EST sequence

5.3 Methods

We devised a scheme as shown in Figure 5.3.

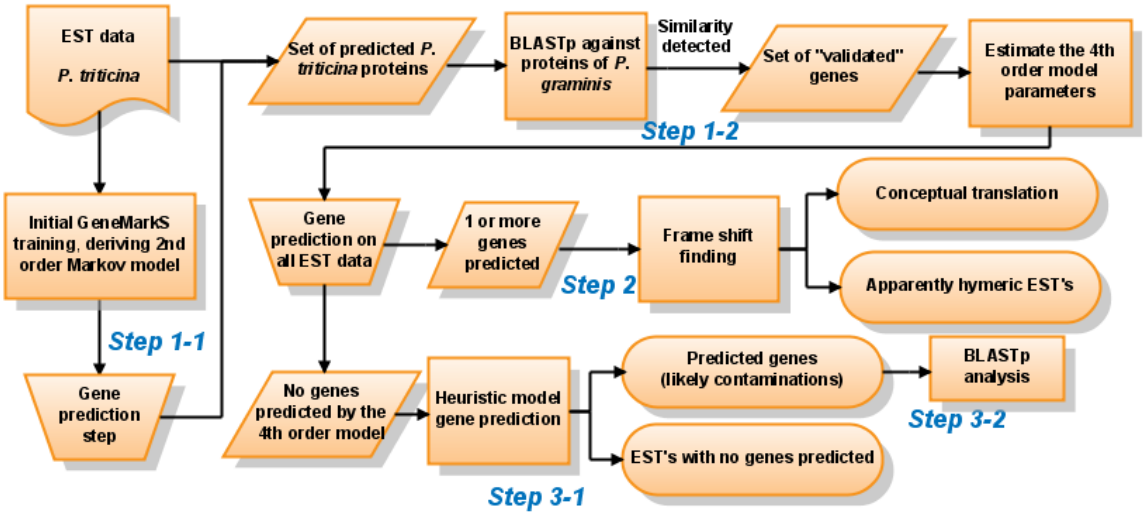


Figure 5.3 Flow chart of Project *Puccinia triticina* sequence analysis

Steps 1-1 and 1-2 are designed to derive a training set for the model for *P. triticina* gene prediction in the EST sequences.

In the initial Step 1-1, all 13,328 EST sequences were concatenated into one single sequence. By running unsupervised training algorithm GeneMarkS (Besemer, Lomsadze et al. 2001), we could derive the parameters of the second order Hidden Markov model. The model was then used in GeneMark.hmm algorithm (Lukashin and Borodovsky 1998) to generate the first set of gene predictions.

Additional Step 1-2 was performed to refine the initial second order model to be a fourth order model. Owing to the fact that these EST could be contaminated by the host genomic sequences, we tried to find out the source of those predicted proteins, either from *P. triticina* or the host. In this procedure, we used proteins evidence from *Puccinia graminis*, a close relative of the wheat leaf rust fungus. We tried to find similarity

between the proteomes of both species by BLASTp. Those predicted genes, which were found to have similarity (on protein level) to proteins of *P. graminis*, were selected to derive a fourth order Markov model. And then this model was applied for gene prediction in all 13,328 sequences.

A sanity check was performed in Step 2 to detect possible frame shift. The EST sequences where several genes were predicted by the fourth order model were further analyzed by GeneTack (Antonov and Borodovsky 2010), a frame shift finding program developed by our group.

Step 3-1 dealt with the EST sequences with zero genes found by the 4^{th} order model, in the event that they could carry genes of the plant host. Our heuristic Markov model (Besemer and Borodovsky 1999) was applied to find plant genes. All predicted genes were used as query to find similarity to the proteins in non-redundant database and to the proteins of *P. graminis* as additional check (Step 3-2).

5.4 Results

5.4.1 GeneMarkS training

The whole set of 13,328 EST fragments were concatenated together for parameter training by GeneMarkS iterations (Besemer, Lomsadze et al. 2001). A second order Hidden Markov model was derived to be used to predict those 10,576 EST fragments which were longer than 300 nucleotides. The short ones are discarded. Figure 5.4 shows the number of gene predictions (0, 1, 2 and more) per EST fragment.

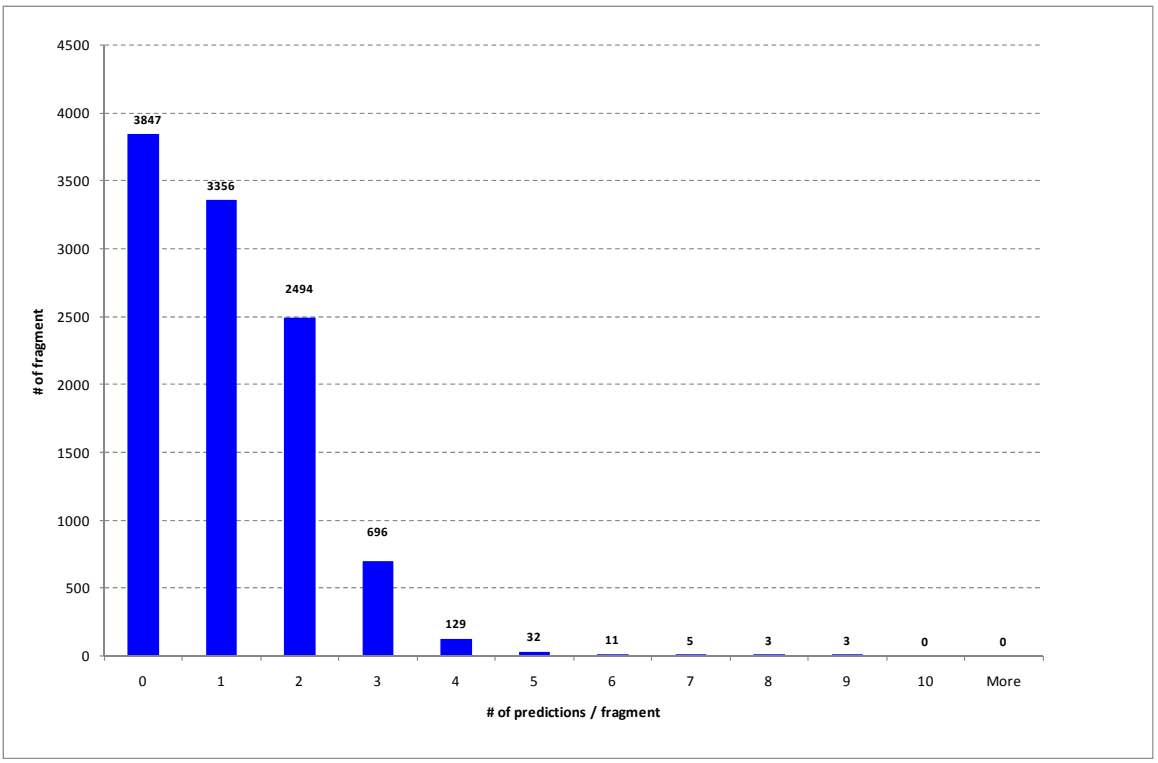


Figure 5.4 Distribution of numbers of EST fragments with 0, 1, 2 … genes predicted by the 2nd order model

Then, we constructed a BLASTp database from all *P. graminis* proteins. We translated those *P. triticina* non-overlapping gene predictions (made by the initial 2nd order model) into proteins, and queried them against the *P. graminis* database. The cut-off e-value was set higher (10^{-10}) than the usual 10^{-5}, since we wanted to make sure to include only those homologous protein into the next training step. This procedure found 2,093 *P. graminis* protein homologs, and the distribution of logarithm values with base of 10 was plotted in Figure 5.5. These low E-value proteins constructed a reliable training set to estimate the parameters of the 4th order Markov model.

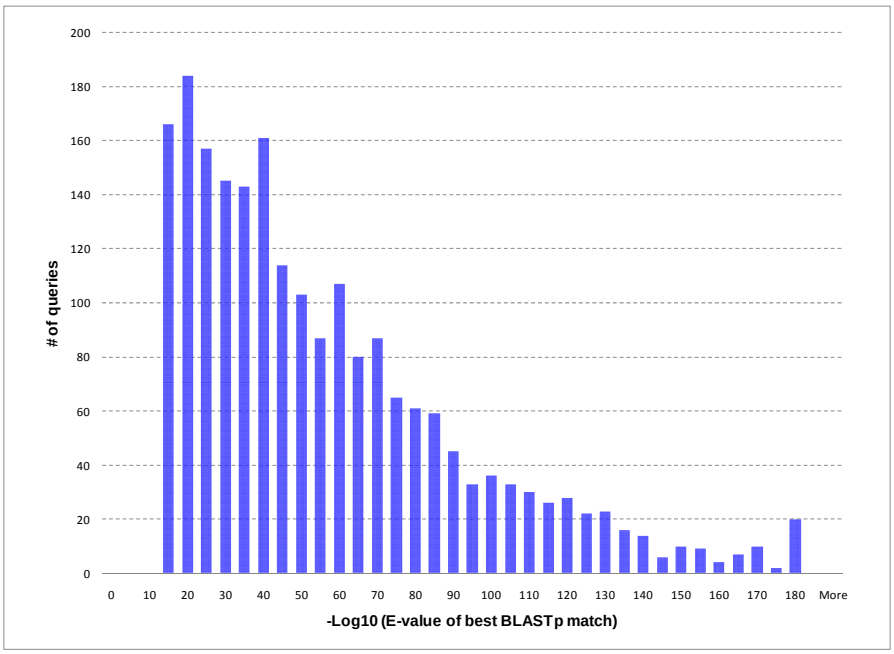


Figure 5.5 Distribution of E-values in BLAST searches between *P. triticina* and *P. graminis* proteins.

Finally, the coding sequences of these 2,093 homologous proteins and the non-coding sequences from the initial gene prediction were used to derive a fourth order Hidden Markov model. Applying this fourth order model, GeneMark.hmm (Lukashin and Borodovsky 1998) was run with this 4th order model on the whole *P. triticina* EST set. Out of 7,594 predictions, 2,519 (33.2%) were found homologous to the *P. graminis* proteins (Table 5.1).

Table 5.1 Gene prediction on the whole EST dataset using the 4th order Markov model

Model	*# of prediction*	*# of hits to PG*	*Percentage*
Order 4	7594	2519	33.2%

It would be interesting to compare the results of gene prediction by the 2nd and 4th order models. Figure 5.6 suggested that there are 5,350 fragments with no gene predicted by the 4th order model, increasing from 3,847 by the 2nd order model. The 4th order model

was more selective, leading to the significant less number of genes predicted. Moreover, the number of fragments with 2 genes predicted dropped to 1,465 from 2,494.

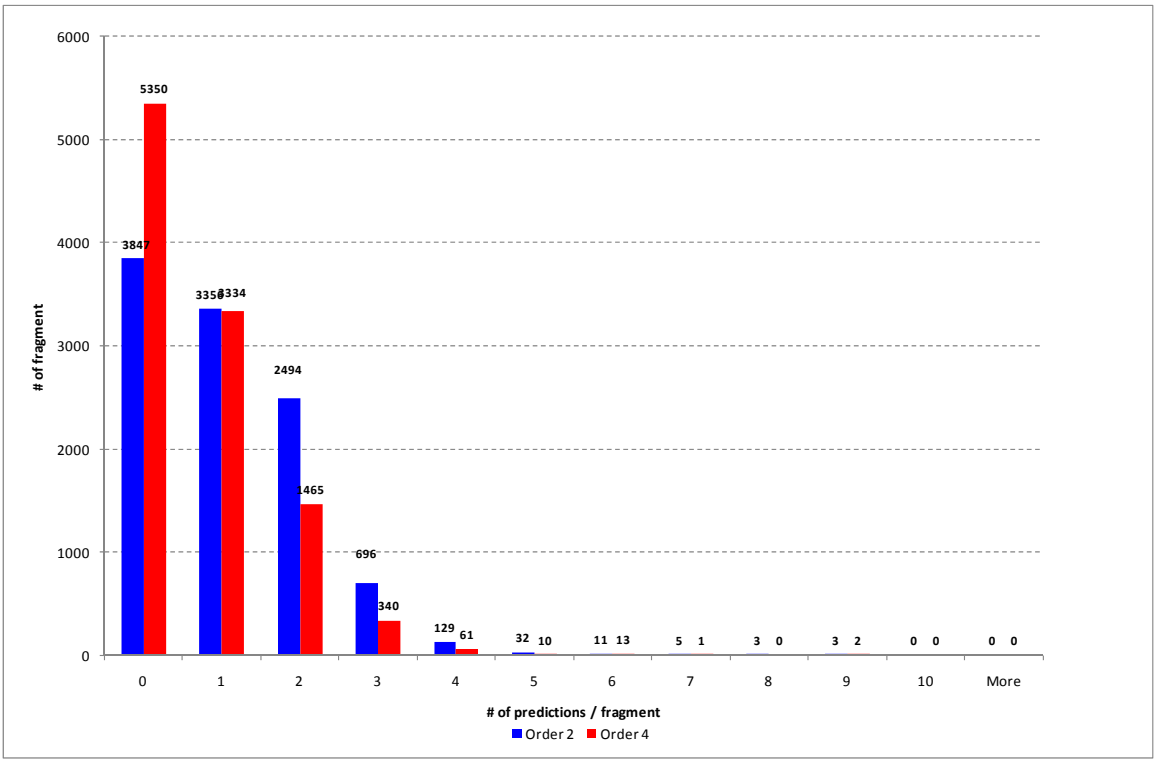


Figure 5.6 Distribution of numbers of EST fragments with 0, 1, 2 … genes predicted by the 2nd order and the 4th order Markov model

A little variation was observed for the distribution of GC composition of the two predictions by the second and fourth order models. This can be explained as follows. The second order model was derived from a larger set of sequences, assuming some possible contamination of the genomic sequences of the host plant (wheat). Thus, the resulting second order model had a property to predict both fungal and plant protein coding regions. The subsequent step, the selection of the homologous genes that had similarity with another fungus, allowed us to build a fungi specific gene model. As a result, this refined model predicted a smaller number of genes which belong to fungal proteome. The GC composition comparison shows that the GC composition of fungal genes is slightly higher than GC composition of the wheat DNA (data not shown).

5.4.2 Detecting frame shifts

Upon application to the EST sequences the GeneMarkS with 4th order model predicted from 0 to 9 coding regions (see Figure 5.6). GeneTack program was developed to detect frame shifts in DNA coding regions. The program (using the 4th order model) was applied to all ESTs containing at least one coding region. In case of several genes predicted in EST we took the ESTs with all genes located on the same strand only (total number = 4,225). We further discarded those sequences with more than 10 consecutive N's or X's, resulting a total of 3,851 ESTs (Table 5.2) to be analyzed by GeneTack. FSdetector predicted no frame shift in 2,645 ESTs (68%) and 1 frame shift in 905 ESTs (24%). Several frame shifts were detected in the rest 301 EST fragments (Table 5.2).

Table 5.2 Frame shift prediction results

Initial data			FSdetector results		
# of predicted genes	Total # of ESTs	# ESTs with all genes located on the same strand	Total # of ESTs analyzed by FSdetector	# of ESTs with no frame shift predicted	# of ESTs with 1 frame shift predicted
1	3334	3334	3018	2302	588
2	1465	768	721	310	279
3	340	111	100	32	32
4	61	8	8	1	4
5	10	2	2	0	1
6	13	2	2	0	1
7	1	0	0	0	0
8	0	0	0	0	0
9	2	0	0	0	0
Total	**5226**	**4225**	**3851**	**2645**	**905**

We further selected only sequences with one predicted frame shift. Conceptual translation was done with respect to predicted frame shift, i.e. first part of a sequence was translated in one frame and second part translated in another frame and then these two parts were joined together. We translated ESTs with 1 or 2 genes predicted only (588+279=867 sequences). Among all conceptually translated protein sequences, only sequences having at least 5 amino acids flanking predicted frame shift (805 sequences total) were used to

validate predictions through BLASTp search against 2 protein databases -- *P. graminis* proteome (20,566 sequences) and NCBI non-redundant database (8,924,078 sequences). E-value threshold 10^{-5} was used for both searches.

Table 5.3 Validation of predicted frame shifts through BLASTp search.

Input data		*P. graminis proteome*			*NCBI non-redundant db*			*Combined results*		
# of gene fragments predicted by GeneMark	*Total # of aa queries*	*# of queries without hits*	*# of TP frame shifts*	*# of potential FP frame*	*# of queries without hits*	*# of TP frame shifts*	*# of potential FP frame*	*# of queries without hits*	*# of TP frame shifts*	*# of potential FP frame shifts*
1	543	286	96	161	333	146	64	226	191	126
2	262	112	84	66	136	91	35	73	134	55
Total	**805**	**398**	**180**	**227**	**469**	**237**	**99**	**299**	**325**	**181**

Hits from searches in both databases were combined (Column "Combined results" in Table 5.3). We looked for hits covering the predicted frame shift. There were 325 frame shift predictions confirmed by this database search. We called those query proteins with hits that did not cover the predicted frame shift region, as "Potential False Positive frame shifts" (181 cases). No hits were found for the rest 299 query proteins.

5.4.3 Test possible contaminations

In Figure 5.6, there was a large set (5,350) of EST fragments without any gene predicted by the *P. triticina* specific 4th order Hidden Markov model. We call this "zero set". Additional analysis could shed some light on the sources, either from the fungus itself or the host genome. Our heuristic model was known as a universal model for gene finding across species (Besemer and Borodovsky 1999). It is capable of identifying genes from

other species (in the same GC content range) and did find 5,156 protein coding genes, tabulated in Table 5.4.

Table 5.4 Number of predictions on each fragment.

# of genes predicted	*1*	*2*	*≥3*	*0*	*Total*
# of EST fragments	2299	1060	234	1757	5350
# of predictions	2299	2120	737	/	5156
# of predictions with length <150nt	1436	1484	581	/	3501

We double checked by BLASTp similarity search against close relative *P. graminis* proteome and the NCBI-nr database, resulting positives of 14 (Table 5.5) and 19 (Table 5.6) proteins, respectively. 9 proteins were found similarity to plants proteins, including Zea mays and rice, indicating the possible contamination. On the other hand, the large number of "no hit" proteins can be explained by the short length of the predicted proteins (Figure 5.7). Of all the 5156 proteins predicted, 3,501 genes were short (less than 150 nucleotides).

Table 5.5 BLASTp analysis of prediction in the zero set, against proteins of *P. graminis* genome

Contig ID	Product	E-value	Score	% Identity
>39_Contig2908	| PGTG_18451 | Puccinia graminis f. sp. tritici GDP-L-fucose synthetase (Red cell NADP(H)-binding protein) (transl	7.0E-53	202.0	73.8
>169_PTDH.cn434.na.ptih	| PGTG_12283 | Puccinia graminis f. sp. tritici phenylalanine ammonia-lyase (translation) (691 aa)	2.0E-51	196.0	80.5
>12444_PT0305.K01.CPTR.ptp	| PGTG_10450 | Puccinia graminis f. sp. tritici hypothetical protein similar to proteasome subunit 1 (translation) (220	2.0E-26	112.0	66.7
>7912_PTDG.cn601.na.ptg	| PGTG_02542 | Puccinia graminis f. sp. tritici 50S ribosomal protein L6 (translation) (289 aa)	2.0E-23	102.0	96.1
>105_PTDG.cn708.na.ptg	| PGTG_02033 | Puccinia graminis f. sp. tritici predicted protein (translation) (396 aa)	7.0E-23	100.0	93.8
>12170_PT0317.M03.C21.ptt	| PGTG_15257 | Puccinia graminis f. sp. tritici hypothetical protein similar to short chain dehydrogenase/reductase f	7.0E-18	84.3	95.0
>6773_Contig7061	| PGTG_02586 | Puccinia graminis f. sp. tritici heat shock protein 82 (translation) (709 aa)	9.0E-18	84.0	81.5
>8442_PTDH.cn765.na.ptih	| PGTG_14765 | Puccinia graminis f. sp. tritici hypothetical protein similar to reverse transcriptase/ribonuclease H (tr	1.0E-15	77.0	60.7
>8302_PTDH.cn399.na.ptih	| PGTG_18751 | Puccinia graminis f. sp. tritici predicted protein (translation) (425 aa)	2.0E-15	76.6	66.0
>10538_PT0132d.A04.BR.pth.ch	| PGTG_03551 | Puccinia graminis f. sp. tritici hypothetical protein (translation) (38 aa)	2.0E-14	73.2	100.0
>11097_PT0281.K07.C21.ptm	| PGTG_14854 | Puccinia graminis f. sp. tritici predicted protein (translation) (227 aa)	1.0E-09	57.8	30.6
>136_PTDG_P007.I04.f1.ptg	| PGTG_07714 | Puccinia graminis f. sp. tritici predicted protein (translation) (505 aa)	5.0E-09	55.1	83.3
>12350_PT0311.L16.C21.ptt	| PGTG_08018 | Puccinia graminis f. sp. tritici predicted protein (translation) (677 aa)	3.0E-08	52.4	35.3
>10100_PT0061d.B10.B7.ptg.chir	| PGTG_13788 | Puccinia graminis f. sp. tritici hypothetical protein (translation) (68 aa)	6.0E-07	48.1	42.4

Table 5.6 BLASTp analysis of prediction in the zero set, against proteins of NCBI non-redundant database

Contig ID	Product	E-value	Score	% Identity
>8303_PTDH.cn400.na.ptih	predicted protein [Escherichia coli str. K-12 substr. MG1655] ┌ref|AP_001785.1| hypothetical protein [Escherichia coli str. K-12 substr. W3110] ┌re	3.0E-55	216.0	100.0
>39_Contig2908	predicted protein [Laccaria bicolor S238N-H82] ┌gb|EDR10406.1| predicted protein [Laccaria bicolor S238N-H82]	2.0E-49	198.0	65.4
>8362_PTDH.cn559.na.ptih	hypothetical protein [Escherichia coli str. K-12 substr. W3110] ┌ref|NP_417326.2| predicted protein [Escherichia coli str. K-12 substr. MG1655] ┌r	1.0E-47	191.0	84.7
>7743_Contig8095	hypothetical protein OrniCp118 [Oryza nivara] ┌ref|YP_052841.1| hypothetical protein OrniCp116 [Oryza nivara] ┌dbj|BAD26820.1| unnamed pr	5.0E-46	186.0	91.1
>13151_TaLr.1164G04.R.pti	AT hook motif-containing protein, putative [Oryza sativa (japonica cultivar-group)]	6.0E-38	159.0	60.5
>12666_PT03339.A02.S6Wu.	ribosomal protein S1 [Sorghum bicolor] ┌gb|ABI60889.1| ribosomal protein S1 [Sorghum bicolor]	7.0E-32	139.0	98.6
>7789_Contig8155	cytochrome b [Puccinia triticina] ┌gb|ABB54705.1| cytochrome b [Puccinia triticina]	4.0E-23	110.0	100.0
>12444_PT0305.K01.CPTR.pt	20S proteasome subunit [Laccaria bicolor S238N-H82] ┌gb|EDR12933.1| 20S proteasome subunit [Laccaria bicolor S238N-H82]	4.0E-20	100.0	55.4
>10604_PT0132b.E11.BR.pth	serine incorporator 3 [Oryza sativa (indica cultivar-group)]	5.0E-17	90.1	97.7
>12705_PT0337.F03.S6Wu.p	chloroplast hypothetical protein [Zea mays subsp. mays] ┌gb|AAR91119.1| chloroplast hypothetical protein [Zea mays]	5.0E-13	77.0	94.7
>87_Contig8054	ATP synthase F0 subunit 9 [Verticillium dahliae] ┌gb|ABC60428.1| ATP synthase F0 subunit 9 [Verticillium dahliae]	1.0E-12	75.5	55.6
>7821_Contig8195	cytochrome oxidase I intronic ORF 10	3.0E-12	74.3	54.1
>7912_PTDG.cn601.na.ptg	hypothetical protein UM02772.1 [Ustilago maydis 521] ┌gb|EAK83683.1| hypothetical protein UM02772.1 [Ustilago maydis 521]	4.0E-11	70.9	62.0
>6773_Contig7061	chaperone [Cryptococcus neoformans var. neoformans JEC21] ┌ref|XP_772092.1| hypothetical protein CNBM1380 [Cryptococcus neoformans va	2.0E-10	68.2	63.0
>105_PTDG.cn708.na.ptg	hypothetical protein MPER_04527 [Moniliophthora perniciosa FA553]	5.0E-10	66.6	49.3
>139_PTDG_P008.K04.r1.ptg	predicted protein [Phaeodactylum tricornutum CCAP 1055/1] ┌gb|EEC50834.1| predicted protein [Phaeodactylum tricornutum CCAP 1055/1]	1.0E-09	65.9	27.1
>169_PTDH.cn434.na.ptih	RecName: Full=Phenylalanine ammonia-lyase ┌emb|CAA31486.1| phenylalanine ammonia-lyase [Rhodotorula mucilaginosa]	1.0E-08	62.0	34.8
>90_Contig8175	hCG23632, isoform CRA_c [Homo sapiens]	3.0E-07	57.4	81.3
>12931_TaLr.1134E01.R.pti	hypothetical protein OsI_09431 [Oryza sativa Indica Group]	3.0E-07	57.8	68.1

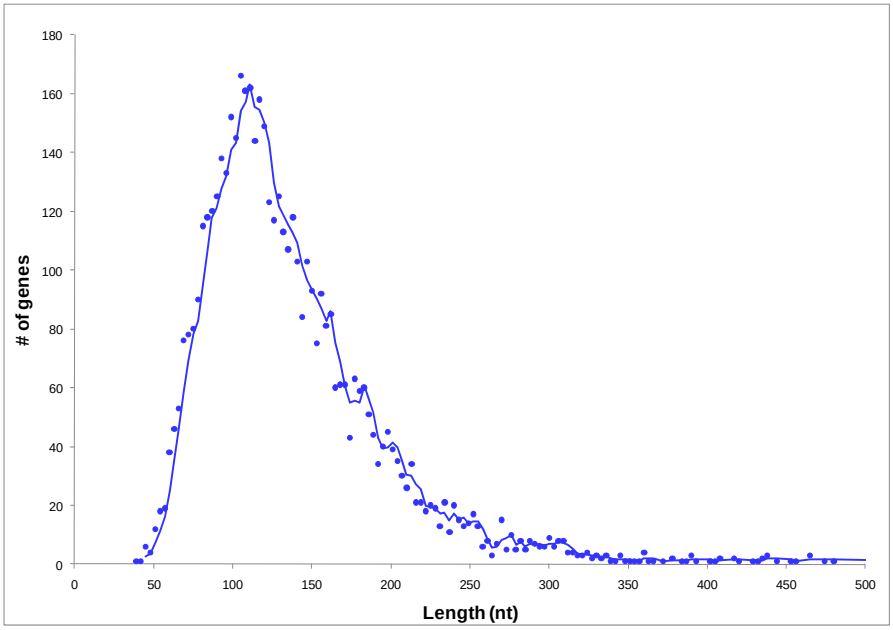


Figure 5.7 Length distribution of genes predicted by heuristic model in zero set.
This set of EST where 4th order model does not predict genes. Most of 5156 predictions are shorter than 150nt.

5.5 Conclusion and data access

This collaboration work was presented in the poster session of 14th congress of the International Society for Molecular Plant-Microbe Interactions in Quebec, Canada (http://www.mpmi2009.ulaval.ca/accueil-mpmi0.html).

All the predictions: gene coordinates, nucleotide sequences of predicted genes as well as translated amino acid sequences are available at:

http://topaz.gatech.edu/~wenhan/collaborators/RustFungal/

The frame shift results are available at:

http://topaz.gatech.edu/~antonov/data/RustFungal/

Dr. Guus Bakkeren's website:

http://www.botany.ubc.ca/people/bakkeren.html

* This chapter was the preliminary result for the following publication in preparation (Bakkeren, Zhu et al. 2010) :

Bakkeren G., Zhu W., Antonov I. and Borodovsky M.

Gene prediction in *Puccinia triticina* based on EST data.

In Preparation.

CHAPTER 6 Conclusions

This dissertation made the effort to improve the *ab initio* gene finding in the new type of metagenomic sequencing data, as well as touching on the annotation result presentation to the general public. The goal was to develop an friendly interface for the wet lab biology researchers to use and design further experiments to clarify interesting biological problems. Recently, an area of research known as "data mining", a branch of a more general topic called "machine learning", has arisen in response to the growing repositories of data of all types. Data mining emphasizes the discovery of new knowledge by finding the inherent patterns in the data. In the case of genomics, the vast amount of DNA sequencing data made it possible for researchers to improve gene finding algorithms. *Ab initio* gene finding methods are statistical approaches to finding genes rely on detecting the subtle statistical variations between coding and non-coding regions and one of the features in the protein coding regions is the bias in the codon usage.

This work explored the codon frequencies in a set of 840 microbial genomes which were completely sequenced and available in 2006. The result showed that there are distinct differences in codon usage, among bacterial and archaeal species, as well as in the other division of mesophilic and thermophilic species. Patterns of dependency of codon frequencies on the genomic nucleotide frequencies were observed. Thus, a reliable inference of codon usage could be drawn from a simple measure, the GC content (guanine and cytosine fraction). This is especially useful for gene finding in very short sequences, such as those several hundred or so ones from the metagenomic sequencing data.

A comprehensive study (Chapter 2) was performed and gene prediction accuracy on metagenomics sequence data was shown to improve. We revisited the heuristic approach to deriving models for gene finding (Besemer and Borodovsky 1999), proposed in 1999 when there were only 17 genomes available. We confirmed the 1999 model gives similar gene prediction accuracy comparing to the model derived from 840 genomes by the same approach. Further, we tried least square fitting methods, polynomial and logistic regression, to fit codon, triplet, tetramer, pentamer and hexamer frequencies of protein coding regions. An average of 96% accuracy was achieved in a test set composed of 50 microbial genomes. Moreover, we tested our new method on real metagenomics data set, seven human and mouse gut microbiomes. The result showed that several thousand more genes could be added into the current annotation. This newly improved method was made available to the world, in the form of both a webpage interface and downloadable program package. The program is fast enough to give gene prediction for large scale metagenomics data. For example, it takes only one and a half minutes to process the sequences from the environment sample of the Sargasso sea (Venter, Remington et al. 2004), which constitutes 1.045 billion nucleotides. This gives researchers the capability in much expediting the process of biological function analysis on future metagenomics data set.

In the case of complete genomes, this dissertation served as a preliminary study for an upgrade of the current GeneMarkS program development (Chapter 3). Several genomic components, such as tRNA, rRNA, protein coding genes and non-coding regions, were analyzed. The signal strength of the upstream regions of translation initiation sites (TIS) was quantified and results showed that they could be used to

pinpoint the exact location of TIS. We successfully applied the GeneMarkS program on the EST data from fungus pathogen *Puccinia triticina* (Chapter 5). We validated the protein product of a subset of the initial predicted genes by finding homology to the proteome of its close relative, *Puccinia graminis*. With another round of model estimation and prediction, we separated the native protein coding genes from the possible contamination of *P. tricina*'s host genome. This application was an example of how *in silico* Bioinformatics could help to reduce the amount of work that the wet lab experiment researchers have to perform in order to arrive a conclusion.

Last but not least, we used new type of sequencing technology (SoLiD) to measure the gene expression level in *Bacillus anthracis* genome (Chapter 4). We further applied this new knowledge to assess the correlation of gene expression level with RBS scores and codon usage bias, in terms of codon adaptation index (CAI) and a new measure average translation speed (ATS).

Overall, more sequencing data give us the opportunities to develop novel and more accurate methods. Future strategies should be devoted to combining the intrinsic and extrinsic evidences in addition to utilizing the complementary strengths of different methods to bring prokaryotic and metagenomic gene prediction further close to a point of complete solution.

APPENDIX

Supplementary Tables

Supplementary Table 1 List of 357 genomes which RefSeq annotated protein-coding regions were used for computing genome wide codon frequencies.
See: http://exon.gatech.edu/GeneMark/metagenome/Training/.

Supplementary Table 2 Fifty prokaryotic species whose genomic sequences were used in the tests (34 bacteria and 16 archaea).

Accession	Species	Kingdom	GC content%	Genome size (nt)	Optimal growth temperature
NC_000854	*Aeropyrum_pernix*	Archaea	56.3	1669696	90-95C
NC_000917	*Archaeoglobus_fulgidus*	Archaea	48.6	2178400	83C
NC_006396	*Haloarcula_marismortui_ATCC_43049*	Archaea	62.4	3131724	40-50C
NC_002607	*Halobacterium_sp*	Archaea	67.9	2014239	42C
NC_000916	*Methanobacterium_thermoautotrophicum*	Archaea	49.5	1751377	65-70C
NC_000909	*Methanococcus_jannaschii*	Archaea	31.4	1664970	85C
NC_003551	*Methanopyrus_kandleri*	Archaea	61.2	1694969	98C
NC_003552	*Methanosarcina_acetivorans*	Archaea	42.7	5751492	35-40C
NC_007681	*Methanosphaera_stadtmanae*	Archaea	27.6	1767403	36-40C
NC_005877	*Picrophilus_torridus_DSM_9790*	Archaea	36.0	1545895	60C
NC_003364	*Pyrobaculum_aerophilum*	Archaea	51.4	2222430	100C
NC_000961	*Pyrococcus_horikoshii*	Archaea	41.9	1738505	98C
NC_003106	*Sulfolobus_tokodaii*	Archaea	32.8	2694756	80C
NC_006624	*Thermococcus_kodakaraensis_KOD1*	Archaea	52.0	2088737	85C
NC_002578	*Thermoplasma_acidophilum*	Archaea	46.0	1564906	59C
NC_002689	*Thermoplasma_volcanium*	Archaea	39.9	1584804	60C
NC_003062	*Agrobacterium_tumefaciens_C58_Cereon*	Bacteria	59.4	2841580	25-28C
NC_003063	*Agrobacterium_tumefaciens_C58_Cereon*	Bacteria	59.3	2075577	25-28C
NC_000918	*Aquifex_aeolicus*	Bacteria	43.5	1551335	96C
NC_000964	*Bacillus_subtilis*	Bacteria	43.5	4214630	25-35C
NC_001318	*Borrelia_burgdorferi*	Bacteria	28.6	910724	N/A
NC_003317	*Brucella_melitensis*	Bacteria	57.2	2117144	37C
NC_003318	*Brucella_melitensis*	Bacteria	57.3	1177787	37C
NC_002163	*Campylobacter_jejuni*	Bacteria	30.5	1641481	N/A
NC_002696	*Caulobacter_crescentus*	Bacteria	67.2	4016947	35C
NC_003030	*Clostridium_acetobutylicum*	Bacteria	30.9	3940880	10-65C
NC_003366	*Clostridium_perfringens*	Bacteria	28.6	3031430	37C
NC_000913	*Escherichia_coli_K12*	Bacteria	50.8	4639675	37C
NC_000907	*Haemophilus_influenzae*	Bacteria	38.2	1830138	35-37C
NC_000915	*Helicobacter_pylori_26695*	Bacteria	38.9	1667867	37C
NC_002678	*Mesorhizobium_loti*	Bacteria	62.7	7036071	N/A
NC_006361	*Nocardia_farcinica_IFM10152*	Bacteria	70.8	6021225	37C
NC_002663	*Pasteurella_multocida*	Bacteria	40.4	2257487	37C
NC_002516	*Pseudomonas_aeruginosa*	Bacteria	66.6	6264404	25-30C
NC_004578	*Pseudomonas_syringae_tomato_DC3000*	Bacteria	58.4	6397126	N/A

NC_003295	*Ralstonia_solanacearum*	Bacteria	67.0	3716413	N/A
NC_003296	*Ralstonia_solanacearum*	Bacteria	66.9	2094509	N/A
NC_003198	*Salmonella_typhi*	Bacteria	52.1	4809037	37C
NC_003197	*Salmonella_typhimurium_LT2*	Bacteria	52.2	4857432	37C
NC_003037	*Sinorhizobium_meliloti*	Bacteria	60.4	1354226	25-30C
NC_003047	*Sinorhizobium_meliloti*	Bacteria	62.7	3654135	25-30C
NC_003078	*Sinorhizobium_meliloti*	Bacteria	62.4	1683333	25-30C
NC_002758	*Staphylococcus_aureus_Mu50*	Bacteria	32.9	2878529	30-37C
NC_002737	*Streptococcus_pyogenes_M1_GAS*	Bacteria	38.5	1852441	30-35C
NC_003888	*Streptomyces_coelicolor*	Bacteria	72.1	8667507	25-35C
NC_000911	*Synechocystis_PCC6803*	Bacteria	47.7	3573470	N/A
NC_000853	*Thermotoga_maritima*	Bacteria	46.2	1860725	80C
NC_004603	*Vibrio_parahaemolyticus*	Bacteria	45.4	3288558	20-30C
NC_004605	*Vibrio_parahaemolyticus*	Bacteria	45.4	1877212	20-30C
NC_003919	*Xanthomonas_citri*	Bacteria	64.8	5175554	25-30C

Supplementary Table 3 Accuracy of gene prediction in 700nt long fragments from 50 genomic sequences by MetaGene, MetaGeneAnnotator and GeneMark.hmm (GM.hmm) with the heuristic models HAL-99, C-3BA, C-3MT (dn =100 and dc =800).

Accession	Species	Kingdom	GC content%	Optimal growth temperature	Number of annotated complete and partial genes	MetaGene		MetaGene Annotator		GM.hmm with HAL-99		GM.hmm with C-3BA model		GM.hmm with C-3MT model	
						Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp
NC_000854	*Aeropyrum_pernix*	Archaea	56.3	90-95C	1455	97.73	94.61	97.80	95.44	94.71	97.25	94.57	97.66	95.67	97.96
NC_000917	*Archaeoglobus_fulgidus*	Archaea	48.6	83C	1883	98.57	94.94	98.73	95.28	96.60	94.49	98.09	95.01	97.93	94.95
NC_006396	*Haloarcula_marismortui_ATCC_43049*	Archaea	62.4	40-50C	3081	98.41	91.27	97.73	92.42	97.11	94.80	97.27	94.84	96.92	94.85
NC_002607	*Halobacterium_sp*	Archaea	67.9	42C	1346	98.89	91.60	98.51	92.92	98.51	95.12	99.03	95.35	98.96	95.62
NC_000916	*Methanobacterium_thermoautotrophicum*	Archaea	49.5	65-70C	1586	97.86	94.69	98.36	94.89	94.20	95.95	97.98	96.58	97.41	96.38
NC_000909	*Methanococcus_jannaschii*	Archaea	31.4	85C	1297	98.61	93.84	98.61	92.95	96.84	95.51	97.53	95.40	97.69	95.41
NC_003551	*Methanopyrus_kandleri*	Archaea	61.2	98C	1841	94.73	87.81	95.49	89.92	92.99	92.39	93.75	92.70	93.43	93.02
NC_003552	*Methanosarcina_acetivorans*	Archaea	42.7	35-40C	4222	94.24	82.65	94.29	83.67	95.05	85.57	96.00	84.65	95.93	85.23
NC_007681	*Methanosphaera_stadtmanae*	Archaea	27.6	36-40C	1765	99.43	92.37	99.43	93.30	97.90	96.64	98.98	96.73	99.09	97.06
NC_005877	*Picrophilus_torridus_DSM_9790*	Archaea	36.0	60C	2193	98.50	94.86	98.63	96.22	93.11	95.78	96.81	96.98	96.35	97.15
NC_003364	*Pyrobaculum_aerophilum*	Archaea	51.4	100C	1711	92.29	93.27	93.05	93.37	92.29	95.81	93.28	95.63	93.45	95.46
NC_000961	*Pyrococcus_horikoshii*	Archaea	41.9	98C	1001	98.20	96.56	98.30	96.66	97.30	97.99	98.60	97.72	98.80	97.15
NC_003106	*Sulfolobus_tokodaii*	Archaea	32.8	80C	1330	97.97	86.98	97.67	86.31	97.37	91.07	97.59	90.52	98.12	90.50
NC_006624	*Thermococcus_kodakaraensis_KOD1*	Archaea	52.0	85C	2229	98.38	96.95	98.21	96.94	96.55	96.16	97.80	97.06	97.40	96.70
NC_002578	*Thermoplasma_acidophilum*	Archaea	46.0	59C	1484	97.30	92.56	96.97	93.44	94.41	93.09	96.97	93.08	96.77	93.25
NC_002689	*Thermoplasma_volcanium*	Archaea	39.9	60C	2132	94.42	89.63	94.23	90.41	93.06	93.45	95.31	92.87	95.17	93.63
NC_003062	*Agrobacterium_tumefaciens_C58_Cereon chromosome circular*	Bacteria	59.4	25-28C	3274	98.72	94.28	98.35	96.64	94.11	93.82	97.83	96.74	97.83	96.13
NC_003063	*Agrobacterium_tumefaciens_C58_Cereon chromosome linear*	Bacteria	59.3	25-28C	2710	98.04	94.22	97.71	96.26	93.58	94.10	96.68	96.50	96.94	96.23
NC_000918	*Aquifex_aeolicus*	Bacteria	43.5	96C	1491	97.38	91.21	97.45	91.44	96.24	92.22	97.85	92.05	97.72	91.87
NC_000964	*Bacillus_subtilis*	Bacteria	43.5	25-35C	3034	96.34	93.06	96.51	93.61	95.52	95.74	96.57	95.50	96.54	95.44
NC_001318	*Borrelia_burgdorferi*	Bacteria	28.6	N/A	961	97.92	95.44	98.23	95.93	95.63	97.04	96.77	97.18	96.46	97.27
NC_003317	*Brucella_melitensis*	Bacteria	57.2	37C	3201	96.44	91.98	96.38	93.88	87.94	87.45	94.97	93.17	94.88	92.65
NC_003318	*Brucella_melitensis*	Bacteria	57.3	37C	1959	95.71	92.18	95.56	93.69	88.36	89.55	94.33	93.38	94.79	92.94
NC_002163	*Campylobacter_jejuni*	Bacteria	30.5	N/A	2977	98.42	94.98	98.59	95.29	96.91	96.01	97.88	96.43	97.25	96.40
NC_002696	*Caulobacter_crescentus*	Bacteria	67.2	35C	4115	98.96	93.95	98.71	95.78	97.40	95.63	98.35	97.07	98.59	96.85
NC_003030	*Clostridium_acetobutylicum*	Bacteria	30.9	10-65C	5050	98.44	92.71	98.69	94.00	96.83	96.03	97.39	96.09	97.21	96.22
NC_003366	*Clostridium_perfringens*	Bacteria	28.6	37C	3305	99.30	93.80	99.46	94.56	97.55	97.31	98.97	98.14	98.79	98.05
NC_000913	*Escherichia_coli_K12*	Bacteria	50.8	37C	8995	96.41	91.45	95.94	93.18	90.48	91.96	95.40	93.68	95.48	93.38
NC_000907	*Haemophilus_influenzae*	Bacteria	38.2	35-37C	2421	98.88	90.44	98.64	91.18	95.46	92.51	97.81	93.01	97.69	92.89
NC_000915	*Helicobacter_pylori_26695*	Bacteria	38.9	37C	1798	96.44	93.78	96.50	94.04	94.94	95.58	96.72	96.08	96.44	95.85
NC_002678	*Mesorhizobium_loti*	Bacteria	62.7	N/A	6643	97.59	90.46	97.49	92.85	95.57	93.11	96.94	94.57	97.17	94.25
NC_006361	*Nocardia_farcinica_IFM10152*	Bacteria	70.8	37C	5709	99.02	92.87	98.77	94.98	96.65	96.84	97.55	97.51	97.74	97.54
NC_002663	*Pasteurella_multocida*	Bacteria	40.4	37C	1856	99.03	95.33	99.14	96.54	95.15	97.14	98.44	97.81	98.38	98.07
NC_002516	*Pseudomonas_aeruginosa*	Bacteria	66.6	25-30C	6326	99.07	93.86	98.99	95.72	96.90	94.89	98.42	96.99	98.69	96.60
NC_004578	*Pseudomonas_syringae_tomato_DC3000*	Bacteria	58.4	N/A	7998	97.90	89.53	97.70	91.54	92.29	88.60	96.60	92.31	96.77	91.97
NC_003295	*Ralstonia_solanacearum*	Bacteria	67.0	N/A	4376	98.40	91.71	98.54	94.27	96.62	95.66	97.67	96.74	98.01	96.38
NC_003296	*Ralstonia_solanacearum*	Bacteria	66.9	N/A	2350	97.70	89.86	97.45	92.68	95.57	94.73	96.89	95.39	96.98	95.04
NC_003198	*Salmonella_typhi*	Bacteria	52.1	37C	6253	96.95	88.21	96.71	89.52	92.55	89.36	95.60	90.13	95.86	89.81
NC_003197	*Salmonella_typhimurium_LT2*	Bacteria	52.2	37C	8376	96.18	92.74	95.71	94.32	90.93	93.48	94.87	94.67	95.13	94.51
NC_003047	*Sinorhizobium_meliloti_1021*	Bacteria	62.7	25-30C	4012	98.80	93.20	98.58	95.60	96.56	95.25	97.43	96.35	97.51	95.81
NC_003037	*Sinorhizobium_meliloti_1021_plasmid_pSymA*	Bacteria	60.4	25-30C	1557	95.57	88.47	95.50	90.45	93.06	90.68	93.96	92.42	94.03	91.50
NC_003078	*Sinorhizobium_meliloti_1021_plasmid_pSymB*	Bacteria	62.4	25-30C	2028	97.73	92.44	97.53	94.78	95.02	94.79	96.75	96.32	96.89	95.71
NC_002758	*Staphylococcus_aureus_Mu50*	Bacteria	32.9	30-37C	3185	98.90	92.78	99.00	94.29	96.95	97.05	97.71	96.98	97.33	97.18
NC_002737	*Streptococcus_pyogenes_M1_GAS*	Bacteria	38.5	30-35C	2120	98.87	91.17	98.44	90.86	95.71	92.90	96.93	93.71	96.98	93.41
NC_003888	*Streptomyces_coelicolor*	Bacteria	72.1	25-35C	10131	98.50	92.40	98.47	94.72	96.56	96.68	96.90	97.68	97.53	97.52
NC_000911	*Synechocystis_PCC6803*	Bacteria	47.7	N/A	2735	96.27	93.04	95.43	94.05	92.58	93.33	95.14	94.86	95.14	94.65
NC_000853	*Thermotoga_maritima*	Bacteria	46.2	80C	1976	96.15	94.34	96.46	94.73	95.19	95.43	97.47	95.73	97.52	95.59
NC_004603	*Vibrio_parahaemolyticus*	Bacteria	45.4	20-30C	3855	97.69	93.80	97.04	95.14	94.37	96.60	96.42	96.85	96.55	96.85
NC_004605	*Vibrio_parahaemolyticus*	Bacteria	45.4	20-30C	1805	97.34	93.66	96.90	95.31	94.96	96.29	96.57	96.78	96.40	96.67
NC_003919	*Xanthomonas_citri*	Bacteria	64.8	25-30C	5645	97.80	90.27	97.82	93.83	94.49	94.96	96.65	96.70	96.86	96.28
Average of archaea						**97.22**	**92.16**	**97.25**	**92.76**	**95.50**	**94.44**	**96.85**	**94.55**	**96.82**	**94.65**
Average of bacteria						**97.73**	**92.46**	**97.60**	**93.99**	**94.67**	**94.20**	**96.84**	**95.46**	**96.88**	**95.22**
Average of all						**97.57**	**92.36**	**97.49**	**93.60**	**94.93**	**94.28**	**96.84**	**95.17**	**96.86**	**95.04**

Supplementary Table 4 Accuracy of gene prediction in 400nt long fragments from 50 genomic sequences by MetaGene, MetaGeneAnnotator and GeneMark.hmm (GM.hmm) with the heuristic models HAL-99, C-3BA, C-3MT (dn =100 and dc =800).

Accession	Species	Kingdom	GC content%	Optimal growth temperature	Number of annotated complete and partial genes	MetaGene		MetaGene Annotator		GM.hmm with HAL-99		GM.hmm with C-3BA model		GM.hmm with C-3MT model	
						Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp
NC_000854	*Aeropyrum_pernix*	Archaea	56.3	90-95C	2400	97.75	94.18	97.96	94.42	94.50	96.35	93.92	96.99	95.33	96.54
NC_000917	*Archaeoglobus_fulgidus*	Archaea	48.6	83C	3100	98.42	94.72	98.39	94.93	95.97	94.35	97.77	94.48	97.68	94.42
NC_006396	*Haloarcula_marismortui_ATCC_43049*	Archaea	62.4	40-50C	5066	98.22	89.56	97.53	91.21	95.78	94.18	96.01	94.37	95.72	94.41
NC_002607	*Halobacterium_sp*	Archaea	67.9	42C	2338	98.93	90.14	98.67	91.37	98.20	94.37	98.67	95.02	98.63	95.01
NC_000916	*Methanobacterium_thermoautotrophicum*	Archaea	49.5	65-70C	2607	98.12	94.99	98.04	94.60	93.98	95.89	97.70	96.55	97.16	96.72
NC_000909	*Methanococcus_jannaschii*	Archaea	31.4	85C	2195	98.68	92.72	98.41	92.51	96.22	95.05	97.49	95.79	97.77	95.93
NC_003551	*Methanopyrus_kandleri*	Archaea	61.2	98C	2817	94.89	87.90	95.70	90.02	92.33	92.40	93.68	93.09	93.36	93.20
NC_003552	*Methanosarcina_acetivorans*	Archaea	42.7	35-40C	6849	93.44	79.29	93.62	79.90	93.56	83.84	94.67	83.77	94.25	84.08
NC_007681	*Methanosphaera_stadtmanae*	Archaea	27.6	36-40C	2840	99.40	91.86	99.37	92.55	97.57	96.79	98.80	96.89	98.70	97.12
NC_005877	*Picrophilus_torridus_DSM_9790*	Archaea	36.0	60C	3342	97.61	93.36	97.55	94.41	91.38	95.23	95.75	96.88	95.24	97.22
NC_003364	*Pyrobaculum_aerophilum*	Archaea	51.4	100C	2812	93.63	91.90	94.49	92.26	93.53	95.50	94.35	95.57	95.16	95.74
NC_000961	*Pyrococcus_horikoshii*	Archaea	41.9	98C	1730	97.92	95.01	98.09	95.02	96.65	97.21	98.15	96.26	98.32	95.94
NC_003106	*Sulfolobus_tokodaii*	Archaea	32.8	80C	2327	97.51	83.88	97.59	84.36	96.61	90.75	97.16	90.58	97.38	90.53
NC_006624	*Thermococcus_kodakaraensis_KOD1*	Archaea	52.0	85C	3619	98.78	96.57	98.76	97.04	96.77	96.37	98.51	97.09	98.15	96.76
NC_002578	*Thermoplasma_acidophilum*	Archaea	46.0	59C	2381	96.68	90.95	96.64	92.22	93.57	92.56	96.72	93.43	96.26	93.78
NC_002689	*Thermoplasma_volcanium*	Archaea	39.9	60C	3241	93.58	87.84	93.80	88.89	92.50	93.92	95.22	93.20	94.72	93.57
NC_003062	*Agrobacterium_tumefaciens_C58_Cereon chromosome circular*	Bacteria	59.4	25-28C	5240	98.28	92.74	97.65	94.67	90.90	90.83	97.00	95.69	96.97	95.24
NC_003063	*Agrobacterium_tumefaciens_C58_Cereon chromosome linear*	Bacteria	59.3	25-28C	4204	97.69	92.75	97.22	94.87	90.49	90.92	95.96	95.71	96.08	95.46
NC_000918	*Aquifex_aeolicus*	Bacteria	43.5	96C	2440	97.13	91.51	97.01	91.92	95.86	93.30	97.54	92.97	97.87	92.99
NC_000964	*Bacillus_subtilis*	Bacteria	43.5	25-35C	4956	95.48	91.44	95.94	91.92	94.73	95.10	95.88	94.70	95.92	94.87
NC_001318	*Borrelia_burgdorferi*	Bacteria	28.6	N/A	1561	98.08	95.21	98.33	95.76	96.22	97.60	96.86	97.99	96.48	98.24
NC_003317	*Brucella_melitensis chromosome I*	Bacteria	57.2	37C	4843	95.77	90.36	95.02	92.30	83.79	84.86	93.64	92.42	93.60	91.97
NC_003318	*Brucella_melitensis chromosome II*	Bacteria	57.3	37C	2896	94.96	90.34	94.75	92.33	84.46	85.58	93.75	92.60	93.75	92.00
NC_002163	*Campylobacter_jejuni*	Bacteria	30.5	N/A	4367	98.33	94.46	98.49	95.01	96.18	95.82	97.05	96.52	96.27	96.47
NC_002696	*Caulobacter_crescentus*	Bacteria	67.2	35C	6742	98.55	93.28	98.47	95.21	96.48	95.20	97.82	97.06	97.98	96.88
NC_003030	*Clostridium_acetobutylicum*	Bacteria	30.9	10-65C	7766	98.17	91.31	98.40	92.51	96.42	95.75	96.97	96.02	96.72	96.06
NC_003366	*Clostridium_perfringens*	Bacteria	28.6	37C	5226	98.95	92.70	99.23	93.66	97.42	96.86	98.43	97.48	98.39	97.70
NC_000913	*Escherichia_coli_K12*	Bacteria	50.8	37C	12700	95.86	90.14	95.32	91.30	88.06	90.13	94.44	93.04	94.51	92.95
NC_000907	*Haemophilus_influenzae*	Bacteria	38.2	35-37C	3698	98.84	89.10	98.89	90.18	94.97	91.79	97.24	92.56	97.11	92.53
NC_000915	*Helicobacter_pylori_26695*	Bacteria	38.9	37C	2866	95.64	91.98	95.53	93.10	93.96	94.99	95.46	95.90	95.36	96.00
NC_002678	*Mesorhizobium_loti*	Bacteria	62.7	N/A	10844	97.31	88.82	97.10	91.33	94.32	91.64	95.94	93.49	96.41	93.12
NC_006361	*Nocardia_farcinica_IFM10152*	Bacteria	70.8	37C	9405	98.64	91.77	98.62	94.34	96.70	96.58	97.48	97.60	97.72	97.49
NC_002663	*Pasteurella_multocida*	Bacteria	40.4	37C	3017	98.81	94.25	98.77	95.00	94.73	96.20	97.65	97.42	97.78	97.52
NC_002516	*Pseudomonas_aeruginosa*	Bacteria	66.6	25-30C	10213	99.00	93.05	98.94	94.96	96.07	94.20	98.29	96.57	98.60	96.38
NC_004578	*Pseudomonas_syringae_tomato_DC3000*	Bacteria	58.4	N/A	12424	97.24	88.23	97.19	89.99	89.67	86.79	95.50	91.66	95.70	91.39
NC_003295	*Ralstonia_solanacearum GMI1000*	Bacteria	67.0	N/A	6886	98.45	90.66	98.48	93.29	95.80	94.54	97.28	96.19	97.65	95.77
NC_003296	*Ralstonia_solanacearum GMI1000 plasmid pGMI1000MP*	Bacteria	66.9	N/A	3739	97.57	87.88	97.57	91.11	95.03	93.67	96.20	94.91	96.58	94.50
NC_003198	*Salmonella_typhi*	Bacteria	52.1	37C	9511	96.67	86.97	96.26	88.33	89.38	87.50	94.14	89.56	94.62	89.45
NC_003197	*Salmonella_typhimurium_LT2*	Bacteria	52.2	37C	12139	95.82	91.07	95.37	92.69	89.15	92.10	94.16	94.34	94.55	94.39
NC_003047	*Sinorhizobium_meliloti 1021*	Bacteria	62.7	25-30C	6371	98.13	91.35	98.02	93.17	94.77	93.37	96.66	95.44	97.08	94.80
NC_003037	*Sinorhizobium_meliloti 1021 plasmid pSymA*	Bacteria	60.4	25-30C	2422	94.88	85.62	94.18	87.29	90.46	87.85	92.44	90.61	93.31	90.18
NC_003078	*Sinorhizobium_meliloti 1021 plasmid pSymB*	Bacteria	62.4	25-30C	3169	97.32	90.71	96.53	92.70	93.18	93.15	95.52	95.04	96.06	94.95
NC_002758	*Staphylococcus_aureus_Mu50*	Bacteria	32.9	30-37C	5040	98.27	91.43	98.59	92.41	95.52	96.34	96.65	97.05	96.31	97.00
NC_002737	*Streptococcus_pyogenes_M1_GAS*	Bacteria	38.5	30-35C	3298	98.21	89.80	98.21	90.12	95.45	92.53	96.51	92.85	96.42	92.90
NC_003888	*Streptomyces_coelicolor*	Bacteria	72.1	25-35C	16085	98.36	90.99	98.13	93.43	95.67	96.28	96.30	97.56	96.79	97.42
NC_000911	*Synechocystis_PCC6803*	Bacteria	47.7	N/A	4706	94.77	91.43	94.05	92.13	91.63	92.45	94.35	94.53	94.11	94.27
NC_000853	*Thermotoga_maritima*	Bacteria	46.2	80C	3200	95.16	94.13	95.97	94.67	93.84	95.36	96.66	95.91	96.81	95.85
NC_004603	*Vibrio_parahaemolyticus RIMD 2210633 chromosome I*	Bacteria	45.4	20-30C	6071	97.17	92.90	96.89	93.90	93.26	95.87	96.01	96.41	96.16	96.34
NC_004605	*Vibrio_parahaemolyticus RIMD 2210633 chromosome II*	Bacteria	45.4	20-30C	2956	96.35	91.96	96.35	93.75	93.98	95.96	95.91	96.49	95.87	96.39
NC_003919	*Xanthomonas_citri*	Bacteria	64.8	25-30C	9030	97.71	88.66	97.51	92.36	92.76	93.11	95.74	95.84	96.41	95.61
Average of archaea						**97.10**	**90.93**	**97.16**	**91.61**	**94.94**	**94.05**	**96.54**	**94.37**	**96.49**	**94.44**
Average of bacteria						**97.28**	**91.15**	**97.15**	**92.70**	**93.27**	**93.07**	**96.10**	**95.00**	**96.23**	**94.86**
Average of all						**97.22**	**91.08**	**97.15**	**92.35**	**93.81**	**93.38**	**96.24**	**94.80**	**96.32**	**94.72**

Supplementary Table 5 Accuracy of gene prediction in 700nt long fragments from 50 complete genome sequences by GeneMark.hmm with the heuristic models based on triplets, tetramers, pentamers, hexamers, as well as with C-MBA heuristic model (dn =100 and dc =800).

Accession	Species	Kingdom	GC content%	Optimal growth temperature	C-MBA		3-3BA		3-LBA		4-4BA		5-5BA		6-6BA		6-LBA	
					Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp
NC_000854	*Aeropyrum_pernix*	Archaea	56.3	90-95C	96.43	95.77	94.16	97.86	96.36	97.91	92.37	97.53	90.72	97.35	86.60	97.22	97.04	97.99
NC_000917	*Archaeoglobus_fulgidus*	Archaea	48.6	83C	98.14	92.96	98.41	94.30	98.35	94.39	98.46	94.54	98.57	94.50	98.41	94.54	98.35	94.54
NC_006396	*Haloarcula_marismortui_ATCC_43049*	Archaea	62.4	40-50C	97.40	93.43	97.24	94.01	97.24	93.89	96.88	94.55	97.01	94.68	96.75	94.57	97.21	94.60
NC_002607	*Halobacterium_sp*	Archaea	67.9	42C	99.11	94.61	99.03	95.01	98.81	95.07	99.03	95.42	98.96	95.42	98.89	95.55	99.18	95.22
NC_000916	*Methanobacterium_thermoautotrophicum*	Archaea	49.5	65-70C	98.17	94.65	97.92	95.57	97.73	95.86	98.49	96.36	98.23	96.00	98.42	96.24	98.11	96.17
NC_000909	*Methanococcus_jannaschii*	Archaea	31.4	85C	97.92	93.93	97.61	94.83	98.00	95.06	98.00	95.35	98.30	95.08	97.76	95.48	98.46	95.30
NC_003551	*Methanopyrus_kandleri*	Archaea	61.2	98C	93.10	92.55	93.92	92.86	94.24	92.83	92.02	92.57	92.78	93.18	90.01	93.09	93.21	93.01
NC_003552	*Methanosarcina_acetivorans*	Archaea	42.7	35-40C	97.30	81.43	96.73	82.39	96.78	82.23	96.54	84.06	96.42	83.58	96.19	84.32	96.54	83.94
NC_007681	*Methanosphaera_stadtmanae*	Archaea	27.6	36-40C	98.30	96.66	97.73	97.18	98.47	96.93	99.15	96.05	99.09	96.58	98.98	96.52	98.70	97.65
NC_005877	*Picrophilus_torridus_DSM_9790*	Archaea	36.0	60C	97.58	95.92	97.58	96.57	98.08	96.46	97.36	97.05	96.72	96.94	96.49	96.93	97.99	96.89
NC_003364	*Pyrobaculum_aerophilum*	Archaea	51.4	100C	93.92	94.47	93.34	95.06	93.28	95.28	93.40	95.23	93.40	95.23	92.75	94.63	94.16	94.99
NC_000961	*Pyrococcus_horikoshii*	Archaea	41.9	98C	98.90	95.38	98.70	97.24	98.70	97.34	98.50	97.53	98.90	97.83	98.70	97.73	98.80	98.02
NC_003106	*Sulfolobus_tokodaii*	Archaea	32.8	80C	98.05	88.83	97.52	90.26	98.35	89.47	97.97	90.80	97.97	90.55	98.12	90.69	98.05	90.18
NC_006624	*Thermococcus_kodakaraensis_KOD1*	Archaea	52.0	85C	98.30	94.03	98.25	96.65	98.16	96.90	98.12	96.68	98.16	96.86	97.94	96.94	98.21	96.73
NC_002578	*Thermoplasma_acidophilum*	Archaea	46.0	59C	97.10	91.09	97.10	92.08	96.97	92.48	97.37	92.81	97.10	92.91	96.90	92.89	97.10	93.27
NC_002689	*Thermoplasma_volcanium*	Archaea	39.9	60C	95.17	91.85	95.40	93.26	95.45	93.05	95.36	93.39	94.79	93.22	94.65	93.69	95.17	93.85
NC_003062	*Agrobacterium_tumefaciens_C58_Cereon chromosome circular*	Bacteria	59.4	25-28C	98.41	95.02	97.53	94.75	97.34	94.99	98.59	95.90	98.78	96.25	98.66	96.27	98.50	95.81
NC_003063	*Agrobacterium_tumefaciens_C58_Cereon chromosome linear*	Bacteria	59.3	25-28C	97.45	95.48	97.08	94.98	96.94	94.97	97.60	96.53	98.01	96.62	97.90	96.75	98.04	96.48
NC_000918	*Aquifex_aeolicus*	Bacteria	43.5	96C	98.12	90.31	97.92	91.82	97.72	91.81	97.65	91.75	97.92	92.29	97.72	92.22	98.19	92.37
NC_000964	*Bacillus_subtilis*	Bacteria	43.5	25-35C	97.10	93.91	96.80	94.68	96.97	94.60	97.17	94.70	96.93	94.87	97.03	95.40	97.36	95.05
NC_001318	*Borrelia_burgdorferi*	Bacteria	28.6	N/A	94.59	96.50	93.96	96.37	95.94	97.26	97.71	96.51	97.61	97.10	98.02	97.52	97.29	97.60
NC_003317	*Brucella_melitensis chromosome I*	Bacteria	57.2	37C	96.06	91.00	94.25	90.06	94.13	90.26	96.16	92.57	96.28	92.83	96.44	92.73	96.34	92.84
NC_003318	*Brucella_melitensis chromosome II*	Bacteria	57.3	37C	95.92	91.35	94.28	90.85	94.13	90.88	95.92	92.88	95.97	92.79	96.17	92.67	95.71	92.87
NC_002163	*Campylobacter_jejuni*	Bacteria	30.5	N/A	95.06	95.25	94.26	96.03	96.31	96.08	98.12	95.99	98.19	96.02	98.29	96.25	98.12	95.99
NC_002696	*Caulobacter_crescentus*	Bacteria	67.2	35C	98.37	95.99	98.10	96.16	98.01	96.00	98.30	97.12	98.42	97.19	98.52	97.33	98.76	96.76
NC_003030	*Clostridium_acetobutylicum*	Bacteria	30.9	10-65C	96.50	95.21	96.10	95.85	97.03	95.95	97.62	95.95	97.72	96.22	97.52	96.27	97.45	96.34
NC_003366	*Clostridium_perfringens*	Bacteria	28.6	37C	97.94	97.35	97.34	97.51	98.21	97.45	99.03	98.11	99.00	97.91	98.79	98.28	98.76	98.17
NC_000913	*Escherichia_coli_K12*	Bacteria	50.8	37C	95.65	91.78	95.04	92.22	94.76	92.22	96.36	92.90	96.66	92.97	96.75	93.10	96.42	92.91
NC_000907	*Haemophilus_influenzae*	Bacteria	38.2	35-37C	97.77	92.03	96.98	92.40	96.90	92.22	97.98	92.37	98.06	92.34	98.14	92.31	98.31	92.36
NC_000915	*Helicobacter_pylori_26695*	Bacteria	38.9	37C	96.77	94.51	96.38	95.64	96.22	95.37	97.33	95.32	97.44	95.01	97.50	94.91	97.27	94.85
NC_002678	*Mesorhizobium_loti*	Bacteria	62.7	N/A	97.50	92.66	97.17	92.12	97.25	92.29	97.73	93.92	97.86	94.07	97.64	94.07	98.04	93.71
NC_006361	*Nocardia_farcinica_IFM10152*	Bacteria	70.8	37C	97.78	96.46	97.27	96.74	97.28	96.78	98.14	97.39	98.21	97.48	98.13	97.36	98.18	97.16
NC_002663	*Pasteurella_multocida*	Bacteria	40.4	37C	98.44	96.87	97.52	96.84	97.41	96.84	98.60	97.55	98.76	97.50	98.55	97.81	98.55	97.49
NC_002516	*Pseudomonas_aeruginosa*	Bacteria	66.6	25-30C	98.77	95.70	98.47	95.85	98.50	95.66	98.85	96.83	98.94	96.95	98.96	96.93	99.05	96.65
NC_004578	*Pseudomonas_syringae_tomato_DC3000*	Bacteria	58.4	N/A	97.07	91.03	96.29	90.53	96.09	90.71	97.37	91.72	97.45	92.00	97.52	91.93	97.49	91.84
NC_003295	*Ralstonia_solanacearum GMI1000*	Bacteria	67.0	N/A	98.19	95.19	97.39	95.28	97.60	95.33	98.40	96.29	98.26	96.15	98.26	96.22	98.65	95.51
NC_003296	*Ralstonia_solanacearum GMI1000 plasmid pGMI1000MP*	Bacteria	66.9	N/A	97.02	93.56	96.60	92.50	96.64	92.62	97.66	94.95	97.57	94.91	97.53	95.22	97.79	94.41
NC_003198	*Salmonella_typhi*	Bacteria	52.1	37C	96.50	88.63	95.79	88.71	95.57	88.81	96.69	89.40	97.07	89.46	97.22	89.54	96.91	89.55
NC_003197	*Salmonella_typhimurium_LT2*	Bacteria	52.2	37C	95.81	93.51	95.31	93.50	95.25	93.53	96.32	94.20	96.32	94.20	96.39	94.19	96.20	94.18
NC_003047	*Sinorhizobium_meliloti 1021*	Bacteria	62.7	25-30C	98.28	94.44	97.63	93.78	97.73	93.83	98.28	95.66	98.31	95.91	98.21	95.65	98.40	95.57
NC_003037	*Sinorhizobium_meliloti 1021 plasmid pSymA*	Bacteria	60.4	25-30C	94.54	90.75	94.28	90.12	94.48	89.97	95.05	92.38	95.18	92.74	94.67	93.06	95.38	92.52
NC_003078	*Sinorhizobium_meliloti 1021 plasmid pSymB*	Bacteria	62.4	25-30C	96.94	94.61	96.75	94.28	96.75	94.06	97.29	95.54	97.04	96.00	96.94	96.14	97.29	95.54
NC_002758	*Staphylococcus_aureus_Mu50*	Bacteria	32.9	30-37C	96.80	96.68	95.51	97.22	96.26	97.33	97.74	97.16	97.36	97.33	97.58	97.52	97.71	97.71
NC_002737	*Streptococcus_pyogenes_M1_GAS*	Bacteria	38.5	30-35C	97.08	92.83	96.32	93.03	96.37	92.95	97.64	93.54	97.78	93.59	97.78	93.42	97.83	93.80
NC_003888	*Streptomyces_coelicolor*	Bacteria	72.1	25-35C	96.15	96.79	95.04	97.29	95.48	97.04	97.07	97.76	97.08	97.94	96.87	98.02	96.58	97.49
NC_000911	*Synechocystis_PCC6803*	Bacteria	47.7	N/A	94.59	94.31	94.48	94.48	93.86	94.24	95.94	94.52	95.87	94.38	95.76	94.75	95.28	94.52
NC_000853	*Thermotoga_maritima*	Bacteria	46.2	80C	97.87	93.84	97.93	95.60	97.62	95.64	97.72	95.69	97.77	95.60	97.77	95.50	97.77	95.36
NC_004603	*Vibrio_parahaemolyticus RIMD 2210633 chromosome I*	Bacteria	45.4	20-30C	96.68	96.65	96.24	96.29	95.93	96.48	96.65	96.63	96.89	96.74	96.86	96.89	96.65	96.98
NC_004605	*Vibrio_parahaemolyticus RIMD 2210633 chromosome II*	Bacteria	45.4	20-30C	96.45	96.29	96.01	96.17	96.07	96.23	96.40	96.51	97.06	96.58	96.84	96.68	96.57	96.62
NC_003919	*Xanthomonas_citri*	Bacteria	64.8	25-30C	97.17	94.41	96.62	94.38	96.65	94.07	97.40	95.50	97.57	95.79	97.52	95.74	97.79	95.21
Average of archaea					**97.18**	**92.97**	**96.92**	**94.07**	**97.19**	**94.07**	**96.81**	**94.37**	**96.70**	**94.37**	**96.10**	**94.44**	**97.27**	**94.52**
Average of bacteria					**96.92**	**94.14**	**96.31**	**94.24**	**96.45**	**94.25**	**97.43**	**95.05**	**97.51**	**95.17**	**97.48**	**95.25**	**97.49**	**95.06**
Average of all					**97.00**	**93.77**	**96.51**	**94.18**	**96.69**	**94.19**	**97.23**	**94.83**	**97.25**	**94.91**	**97.04**	**94.99**	**97.42**	**94.89**

Supplementary Table 6 Accuracy of gene prediction in 400nt long fragments from 50 complete genome sequences by GeneMark.hmm with the heuristic models based on triplets, tetramers, pentamers, hexamers, as well as with C-MBA heuristic model (dn =100 and dc =800).

Accession	Species	Kingdom	GC content%	Optimal growth temperature	C-MBA		3-3BA		3-LBA		4-4BA		5-5BA		6-6BA		6-LBA	
					Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp
NC_000854	*Aeropyrum_pernix*	Archaea	56.3	90-95C	97.08	95.26	92.92	96.79	96.63	96.91	91.13	97.03	89.67	97.11	85.75	96.80	97.13	97.37
NC_000917	*Archaeoglobus_fulgidus*	Archaea	48.6	83C	98.13	92.66	98.29	94.25	98.16	94.21	98.10	94.24	98.10	94.50	98.03	94.41	98.29	94.39
NC_006396	*Haloarcula_marismortui_ATCC_43049*	Archaea	62.4	40-50C	96.64	93.51	96.57	94.19	96.55	94.11	96.01	94.52	96.25	94.46	95.64	94.52	96.90	94.90
NC_002607	*Halobacterium_sp*	Archaea	67.9	42C	98.63	93.85	98.37	94.34	98.55	94.78	98.50	94.77	98.80	94.79	98.72	94.98	98.89	94.75
NC_000916	*Methanobacterium_thermoautotrophicum*	Archaea	49.5	65-70C	98.20	94.57	98.12	95.95	97.89	96.16	98.35	96.32	98.39	96.18	98.20	96.42	98.16	96.71
NC_000909	*Methanococcus_jannaschii*	Archaea	31.4	85C	97.18	94.05	97.27	95.65	98.13	95.31	98.00	95.60	98.09	95.39	97.81	95.68	98.27	95.40
NC_003551	*Methanopyrus_kandleri*	Archaea	61.2	98C	92.90	92.64	93.65	93.15	93.82	92.74	91.55	93.27	91.76	93.19	88.71	93.25	92.69	93.32
NC_003552	*Methanosarcina_acetivorans*	Archaea	42.7	35-40C	96.53	80.18	95.69	80.85	95.98	80.69	95.27	82.52	95.15	83.10	94.77	83.58	96.09	82.78
NC_007681	*Methanosphaera_stadtmanae*	Archaea	27.6	36-40C	97.22	96.98	96.76	98.00	98.17	97.38	99.05	96.67	99.08	96.77	98.98	96.90	98.45	97.80
NC_005877	*Picrophilus_torridus_DSM_9790*	Archaea	36.0	60C	95.99	95.36	96.05	96.57	96.83	96.48	96.02	96.71	96.08	96.95	95.21	96.66	97.16	96.69
NC_003364	*Pyrobaculum_aerophilum*	Archaea	51.4	100C	94.88	94.11	94.17	94.94	94.35	95.12	93.95	95.17	94.74	95.14	93.88	95.24	95.66	95.36
NC_000961	*Pyrococcus_horikoshii*	Archaea	41.9	98C	98.44	94.35	98.09	96.15	98.32	96.10	98.09	96.64	98.32	97.20	97.80	97.07	98.21	97.03
NC_003106	*Sulfolobus_tokodaii*	Archaea	32.8	80C	97.03	88.48	96.52	90.06	97.12	89.65	97.25	90.59	96.95	90.42	96.99	90.72	97.25	90.02
NC_006624	*Thermococcus_kodakaraensis_KOD1*	Archaea	52.0	85C	98.54	94.21	98.65	96.51	98.54	96.64	98.67	96.54	98.59	96.56	98.34	96.71	98.62	96.43
NC_002578	*Thermoplasma_acidophilum*	Archaea	46.0	59C	96.81	91.29	96.77	92.57	96.22	91.82	97.02	92.73	96.93	93.29	96.60	93.16	96.72	92.94
NC_002689	*Thermoplasma_volcanium*	Archaea	39.9	60C	95.28	92.12	95.22	93.26	95.03	92.80	95.16	93.60	94.69	94.00	93.61	94.02	94.79	93.52
NC_003062	*Agrobacterium_tumefaciens_C58_Cereon chromosome circular*	Bacteria	59.4	25-28C	97.52	93.62	96.01	92.82	95.74	92.74	97.86	95.10	97.84	94.87	98.02	95.08	97.98	95.13
NC_003063	*Agrobacterium_tumefaciens_C58_Cereon chromosome linear*	Bacteria	59.3	25-28C	96.79	94.39	95.27	93.93	95.24	94.03	97.24	95.98	97.48	95.93	97.86	96.55	97.62	96.09
NC_000918	*Aquifex_aeolicus*	Bacteria	43.5	96C	98.24	91.24	97.58	92.90	97.79	92.66	97.46	92.96	97.58	93.26	97.54	93.26	97.83	92.99
NC_000964	*Bacillus_subtilis*	Bacteria	43.5	25-35C	96.25	93.29	95.80	93.91	95.84	93.76	96.51	94.51	96.61	95.02	96.31	94.87	96.85	94.71
NC_001318	*Borrelia_burgdorferi*	Bacteria	28.6	N/A	91.48	96.88	91.80	97.42	95.84	98.10	97.50	97.88	97.37	98.06	97.44	98.26	96.93	98.12
NC_003317	*Brucella_melitensis chromosome I*	Bacteria	57.2	37C	95.04	89.88	92.67	88.98	92.09	88.76	95.31	92.03	95.44	92.09	95.54	92.10	95.11	91.88
NC_003318	*Brucella_melitensis chromosome II*	Bacteria	57.3	37C	94.96	90.46	92.82	89.81	92.54	89.66	95.06	92.04	95.48	92.26	95.58	92.64	95.34	92.62
NC_002163	*Campylobacter_jejuni*	Bacteria	30.5	N/A	92.83	94.87	91.87	95.91	95.19	96.16	97.48	96.42	97.85	96.43	97.76	96.45	97.48	96.44
NC_002696	*Caulobacter_crescentus*	Bacteria	67.2	35C	98.03	95.69	97.64	95.95	97.55	95.72	98.13	97.11	98.03	97.08	98.12	97.19	98.37	96.76
NC_003030	*Clostridium_acetobutylicum*	Bacteria	30.9	10-65C	95.30	94.96	94.97	95.97	96.54	95.83	97.18	95.88	97.14	95.83	96.83	96.00	96.97	95.96
NC_003366	*Clostridium_perfringens*	Bacteria	28.6	37C	97.07	96.70	96.56	97.47	97.90	97.39	98.58	97.48	98.62	97.41	98.56	97.69	98.16	97.64
NC_000913	*Escherichia_coli_K12*	Bacteria	50.8	37C	94.97	91.07	94.20	91.48	93.74	91.34	95.59	92.44	95.91	92.45	95.96	92.44	95.72	92.48
NC_000907	*Haemophilus_influenzae*	Bacteria	38.2	35-37C	97.16	91.59	96.24	91.89	96.46	91.60	97.89	92.16	97.84	92.11	98.11	92.36	97.94	92.07
NC_000915	*Helicobacter_pylori_26695*	Bacteria	38.9	37C	95.81	94.66	95.57	95.47	95.71	95.01	96.37	95.34	96.72	95.45	96.82	95.56	97.10	95.24
NC_002678	*Mesorhizobium_loti*	Bacteria	62.7	N/A	96.83	91.62	96.20	91.29	96.27	91.36	97.04	93.07	96.97	93.28	97.02	93.38	97.29	93.03
NC_006361	*Nocardia_farcinica_IFM10152*	Bacteria	70.8	37C	97.55	95.86	96.92	96.71	96.80	97.02	98.15	97.30	98.22	97.35	98.06	97.33	98.05	97.21
NC_002663	*Pasteurella_multocida*	Bacteria	40.4	37C	97.85	96.25	96.85	96.47	96.85	96.31	97.94	97.33	97.98	97.14	97.91	97.20	98.28	97.21
NC_002516	*Pseudomonas_aeruginosa*	Bacteria	66.6	25-30C	98.61	95.28	97.99	95.41	98.06	95.47	98.83	96.57	98.91	96.59	98.89	96.79	99.09	96.46
NC_004578	*Pseudomonas_syringae_tomato_DC3000*	Bacteria	58.4	N/A	96.34	90.13	95.40	89.94	95.18	90.12	96.56	91.25	96.68	91.36	96.64	91.32	96.84	91.47
NC_003295	*Ralstonia_solanacearum GMI1000*	Bacteria	67.0	N/A	97.76	94.11	97.08	94.58	97.33	94.49	98.05	95.42	98.04	95.45	98.04	95.66	98.21	95.17
NC_003296	*Ralstonia_solanacearum GMI1000 plasmid pGMI1000MP*	Bacteria	66.9	N/A	96.71	93.53	95.64	92.43	95.96	92.83	97.35	94.42	97.41	94.62	97.22	94.81	97.75	94.13
NC_003198	*Salmonella_typhi*	Bacteria	52.1	37C	95.22	88.06	94.76	88.55	94.48	88.69	95.90	89.07	96.32	89.19	96.34	89.37	96.27	89.35
NC_003197	*Salmonella_typhimurium_LT2*	Bacteria	52.2	37C	94.94	92.69	94.56	93.27	94.25	93.09	95.53	93.82	95.69	93.79	95.77	93.86	95.77	93.77
NC_003047	*Sinorhizobium_meliloti_1021*	Bacteria	62.7	25-30C	97.32	93.22	96.44	92.87	96.48	92.77	97.41	95.18	97.72	95.46	97.65	95.50	97.91	95.05
NC_003037	*Sinorhizobium_meliloti_1021_plasmid pSymA*	Bacteria	60.4	25-30C	93.56	89.25	93.19	88.79	93.23	88.62	94.10	90.87	93.97	91.19	93.81	91.43	94.34	90.89
NC_003078	*Sinorhizobium_meliloti_1021_plasmid pSymB*	Bacteria	62.4	25-30C	96.40	93.94	95.42	92.87	95.55	93.23	96.43	95.47	96.43	95.62	96.24	95.61	96.62	95.27
NC_002758	*Staphylococcus_aureus_Mu50*	Bacteria	32.9	30-37C	94.98	96.61	93.37	97.13	94.56	96.91	96.27	97.18	96.15	97.13	96.29	97.24	96.43	97.43
NC_002737	*Streptococcus_pyogenes_M1_GAS*	Bacteria	38.5	30-35C	96.57	92.03	95.06	92.37	95.60	92.44	96.82	92.63	96.85	92.53	97.12	92.60	97.12	92.98
NC_003888	*Streptomyces_coelicolor*	Bacteria	72.1	25-35C	95.52	96.64	93.93	97.30	94.25	96.99	96.52	97.70	96.56	97.72	96.34	97.85	95.82	97.43
NC_000911	*Synechocystis_PCC6803*	Bacteria	47.7	N/A	92.95	93.90	93.09	93.75	91.90	93.29	94.99	93.99	94.86	93.98	95.18	94.18	94.09	93.99
NC_000853	*Thermotoga_maritima*	Bacteria	46.2	80C	97.38	94.42	96.97	95.71	96.97	95.71	96.88	95.56	96.63	95.82	96.38	95.69	96.72	95.82
NC_004603	*Vibrio_parahaemolyticus RIMD 2210633 chromosome I*	Bacteria	45.4	20-30C	96.52	95.88	95.42	96.10	94.91	95.99	96.34	96.58	96.67	96.62	96.67	96.82	96.20	96.64
NC_004605	*Vibrio_parahaemolyticus RIMD 2210633 chromosome II*	Bacteria	45.4	20-30C	96.18	95.76	95.67	96.35	95.33	96.08	96.45	96.32	96.52	95.87	96.45	96.12	96.28	96.15
NC_003919	*Xanthomonas_citri*	Bacteria	64.8	25-30C	96.80	93.44	95.87	93.47	96.09	93.67	96.97	94.68	97.19	94.88	97.18	94.94	97.59	94.71
Average of archaea					**96.84**	**92.73**	**96.44**	**93.95**	**96.89**	**93.81**	**96.38**	**94.18**	**96.35**	**94.32**	**95.57**	**94.38**	**97.08**	**94.34**
Average of bacteria					**96.10**	**93.59**	**95.26**	**93.80**	**95.54**	**93.76**	**96.84**	**94.76**	**96.93**	**94.82**	**96.93**	**94.95**	**96.94**	**94.77**
Average of all					**96.34**	**93.31**	**95.64**	**93.85**	**95.97**	**93.77**	**96.70**	**94.57**	**96.75**	**94.66**	**96.49**	**94.77**	**96.99**	**94.63**

Supplementary Table 7 Domain classification (bacterial vs archaeal by the C-3BA model) as well as type classification (mesophilic or thermophilic by the C-3MT model) for 700nt fragments.

Accession	Species	Kingdom	GC content%	Optimal growth temperature	C-3BA			C-3MT		
					Bac type	Arc type	% Arc	Meso type	Thermo type	% Thermo
NC_000854	*Aeropyrum_pernix*	Archaea	56.3	90-95C	10	1399	99.3	3	1418	99.8
NC_000917	*Archaeoglobus_fulgidus*	Archaea	48.6	83C	23	1921	98.8	21	1921	98.9
NC_006396	*Haloarcula_marismortui_ATCC_43049*	Archaea	62.4	40-50C	492	2668	84.4	1051	2097	66.6
NC_002607	*Halobacterium_sp*	Archaea	67.9	42C	292	1106	79.1	951	442	31.7
NC_000916	*Methanobacterium_thermoautotrophicum*	Archaea	49.5	65-70C	25	1584	98.4	28	1575	98.3
NC_000909	*Methanococcus_jannaschii*	Archaea	31.4	85C	79	1247	94.0	113	1215	91.5
NC_003551	*Methanopyrus_kandleri*	Archaea	61.2	98C	77	1785	95.9	59	1790	96.8
NC_003552	*Methanosarcina_acetivorans*	Archaea	42.7	35-40C	782	4006	83.7	812	3940	82.9
NC_007681	*Methanosphaera_stadtmanae*	Archaea	27.6	36-40C	579	1227	67.9	1520	282	15.6
NC_005877	*Picrophilus_torridus_DSM_9790*	Archaea	36.0	60C	27	2162	98.8	79	2096	96.4
NC_003364	*Pyrobaculum_aerophilum*	Archaea	51.4	100C	52	1617	96.9	26	1649	98.4
NC_000961	*Pyrococcus_horikoshii*	Archaea	41.9	98C	5	1005	99.5	4	1014	99.6
NC_003106	*Sulfolobus_tokodaii*	Archaea	32.8	80C	186	1248	87.0	199	1243	86.2
NC_006624	*Thermococcus_kodakaraensis_KOD1*	Archaea	52.0	85C	47	2199	97.9	41	2204	98.2
NC_002578	*Thermoplasma_acidophilum*	Archaea	46.0	59C	29	1517	98.1	27	1513	98.2
NC_002689	*Thermoplasma_volcanium*	Archaea	39.9	60C	52	2136	97.6	73	2094	96.6
NC_003062	*Agrobacterium_tumefaciens_C58_Cereon chromosome cir*	Bacteria	59.4	25-28C	3248	63	1.9	3216	116	3.5
NC_003063	*Agrobacterium_tumefaciens_C58_Cereon chromosome lin*	Bacteria	59.3	25-28C	2654	61	2.2	2616	114	4.2
NC_000918	*Aquifex_aeolicus*	Bacteria	43.5	96C	21	1564	98.7	20	1566	98.7
NC_000964	*Bacillus_subtilis*	Bacteria	43.5	25-35C	2859	209	6.8	2811	258	8.4
NC_001318	*Borrelia_burgdorferi*	Bacteria	28.6	N/A	519	438	45.8	589	364	38.2
NC_003317	*Brucella_melitensis chromosome I*	Bacteria	57.2	37C	3205	58	1.8	3165	113	3.4
NC_003318	*Brucella_melitensis chromosome II*	Bacteria	57.3	37C	1947	32	1.6	1930	68	3.4
NC_002163	*Campylobacter_jejuni*	Bacteria	30.5	N/A	2831	191	6.3	2737	266	8.9
NC_002696	*Caulobacter_crescentus*	Bacteria	67.2	35C	4022	147	3.5	4007	182	4.3
NC_003030	*Clostridium_acetobutylicum*	Bacteria	30.9	10-65C	995	4123	80.6	2148	2954	57.9
NC_003366	*Clostridium_perfringens*	Bacteria	28.6	37C	593	2740	82.2	2081	1249	37.5
NC_000913	*Escherichia_coli_K12*	Bacteria	50.8	37C	8946	214	2.3	8967	230	2.5
NC_000907	*Haemophilus_influenzae*	Bacteria	38.2	35-37C	2520	26	1.0	2505	41	1.6
NC_000915	*Helicobacter_pylori_26695*	Bacteria	38.9	37C	1766	44	2.4	1731	78	4.3
NC_002678	*Mesorhizobium_loti*	Bacteria	62.7	N/A	6506	304	4.5	6388	461	6.7
NC_006361	*Nocardia_farcinica_IFM10152*	Bacteria	70.8	37C	5173	538	9.4	4926	795	13.9
NC_002663	*Pasteurella_multocida*	Bacteria	40.4	37C	1848	20	1.1	1841	21	1.1
NC_002516	*Pseudomonas_aeruginosa*	Bacteria	66.6	25-30C	6298	121	1.9	6136	327	5.1
NC_004578	*Pseudomonas_syringae_tomato_DC3000*	Bacteria	58.4	N/A	8177	193	2.3	7972	444	5.3
NC_003295	*Ralstonia_solanacearum GMI1000*	Bacteria	67.0	N/A	4315	103	2.3	4311	139	3.1
NC_003296	*Ralstonia_solanacearum GMI1000 plasmid pGMI1000MP*	Bacteria	66.9	N/A	2306	81	3.4	2304	94	3.9
NC_003198	*Salmonella_typhi*	Bacteria	52.1	37C	6483	150	2.3	6476	198	3.0
NC_003197	*Salmonella_typhimurium_LT2*	Bacteria	52.2	37C	8185	208	2.5	8162	269	3.2
NC_003047	*Sinorhizobium_meliloti 1021*	Bacteria	62.7	25-30C	3779	278	6.9	3730	353	8.6
NC_003037	*Sinorhizobium_meliloti 1021 plasmid pSymA*	Bacteria	60.4	25-30C	1354	229	14.5	1272	328	20.5
NC_003078	*Sinorhizobium_meliloti 1021 plasmid pSymB*	Bacteria	62.4	25-30C	1856	181	8.9	1794	259	12.6
NC_002758	*Staphylococcus_aureus_Mu50*	Bacteria	32.9	30-37C	3039	170	5.3	2991	199	6.2
NC_002737	*Streptococcus_pyogenes_M1_GAS*	Bacteria	38.5	30-35C	2114	79	3.6	2099	102	4.6
NC_003888	*Streptomyces_coelicolor*	Bacteria	72.1	25-35C	8604	1446	14.4	6343	3789	37.4
NC_000911	*Synechocystis_PCC6803*	Bacteria	47.7	N/A	2626	117	4.3	2607	142	5.2
NC_000853	*Thermotoga_maritima*	Bacteria	46.2	80C	38	1974	98.1	22	1994	98.9
NC_004603	*Vibrio_parahaemolyticus RIMD 2210633 chromosome I*	Bacteria	45.4	20-30C	3783	55	1.4	3765	78	2.0
NC_004605	*Vibrio_parahaemolyticus RIMD 2210633 chromosome II*	Bacteria	45.4	20-30C	1768	33	1.8	1764	36	2.0
NC_003919	*Xanthomonas_citri*	Bacteria	64.8	25-30C	5548	94	1.7	5530	149	2.6

Supplementary Table 8 Domain classification (bacterial vs archaeal by the C-3BA model) as well as type classification (mesophilic or thermophilic by the C-3MT model) for 400nt fragments.

Accession	Species	Kingdom	GC content%	Optimal growth temperature	C-3BA			C-3MT		
					Bac type	Arc type	% Arc	Meso type	Thermo type	% Thermo
NC_000854	*Aeropyrum_pernix*	Archaea	56.3	90-95C	19	2305	99.2	13	2357	99.5
NC_000917	*Archaeoglobus_fulgidus*	Archaea	48.6	83C	45	3163	98.6	40	3167	98.8
NC_006396	*Haloarcula_marismortui_ATCC_43049*	Archaea	62.4	40-50C	936	4218	81.8	1884	3252	63.3
NC_002607	*Halobacterium_sp*	Archaea	67.9	42C	567	1861	76.6	1625	802	33.0
NC_000916	*Methanobacterium_thermoautotrophicum*	Archaea	49.5	65-70C	36	2602	98.6	37	2582	98.6
NC_000909	*Methanococcus_jannaschii*	Archaea	31.4	85C	144	2090	93.6	262	1975	88.3
NC_003551	*Methanopyrus_kandleri*	Archaea	61.2	98C	145	2690	94.9	109	2713	96.1
NC_003552	*Methanosarcina_acetivorans*	Archaea	42.7	35-40C	1477	6263	80.9	1502	6175	80.4
NC_007681	*Methanosphaera_stadtmanae*	Archaea	27.6	36-40C	995	1901	65.6	2370	516	17.9
NC_005877	*Picrophilus_torridus_DSM_9790*	Archaea	36.0	60C	40	3263	98.8	145	3129	95.6
NC_003364	*Pyrobaculum_aerophilum*	Archaea	51.4	100C	123	2653	95.6	62	2733	97.8
NC_000961	*Pyrococcus_horikoshii*	Archaea	41.9	98C	11	1753	99.4	8	1765	99.5
NC_003106	*Sulfolobus_tokodaii*	Archaea	32.8	80C	402	2094	83.9	441	2062	82.4
NC_006624	*Thermococcus_kodakaraensis_KOD1*	Archaea	52.0	85C	57	3615	98.4	54	3617	98.5
NC_002578	*Thermoplasma_acidophilum*	Archaea	46.0	59C	64	2401	97.4	62	2382	97.5
NC_002689	*Thermoplasma_volcanium*	Archaea	39.9	60C	103	3208	96.9	130	3151	96.0
NC_003062	*Agrobacterium_tumefaciens_C58_Cereon chromosome circular*	Bacteria	59.4	25-28C	5181	131	2.5	5101	234	4.4
NC_003063	*Agrobacterium_tumefaciens_C58_Cereon chromosome linear*	Bacteria	59.3	25-28C	4075	140	3.3	4015	216	5.1
NC_000918	*Aquifex_aeolicus*	Bacteria	43.5	96C	37	2523	98.6	34	2534	98.7
NC_000964	*Bacillus_subtilis*	Bacteria	43.5	25-35C	4543	475	9.5	4454	557	11.1
NC_001318	*Borrelia_burgdorferi*	Bacteria	28.6	N/A	813	730	47.3	866	667	43.5
NC_003317	*Brucella_melitensis chromosome I*	Bacteria	57.2	37C	4800	107	2.2	4735	194	3.9
NC_003318	*Brucella_melitensis chromosome II*	Bacteria	57.3	37C	2863	69	2.4	2828	123	4.2
NC_002163	*Campylobacter_jejuni*	Bacteria	30.5	N/A	4073	318	7.2	3888	470	10.8
NC_002696	*Caulobacter_crescentus*	Bacteria	67.2	35C	6506	289	4.3	6464	355	5.2
NC_003030	*Clostridium_acetobutylicum*	Bacteria	30.9	10-65C	1616	6227	79.4	3296	4523	57.8
NC_003366	*Clostridium_perfringens*	Bacteria	28.6	37C	1056	4221	80.0	3124	2139	40.6
NC_000913	*Escherichia_coli_K12*	Bacteria	50.8	37C	12561	330	2.6	12544	370	2.9
NC_000907	*Haemophilus_influenzae*	Bacteria	38.2	35-37C	3835	50	1.3	3787	94	2.4
NC_000915	*Helicobacter_pylori_26695*	Bacteria	38.9	37C	2757	96	3.4	2695	152	5.3
NC_002678	*Mesorhizobium_loti*	Bacteria	62.7	N/A	10498	631	5.7	10343	885	7.9
NC_006361	*Nocardia_farcinica_IFM10152*	Bacteria	70.8	37C	8382	1011	10.8	7873	1555	16.5
NC_002663	*Pasteurella_multocida*	Bacteria	40.4	37C	2992	32	1.1	2992	33	1.1
NC_002516	*Pseudomonas_aeruginosa*	Bacteria	66.6	25-30C	10161	234	2.3	9800	648	6.2
NC_004578	*Pseudomonas_syringae_tomato_DC3000*	Bacteria	58.4	N/A	12596	349	2.7	12201	809	6.2
NC_003295	*Ralstonia_solanacearum GMI1000*	Bacteria	67.0	N/A	6801	163	2.3	6778	243	3.5
NC_003296	*Ralstonia_solanacearum GMI1000 plasmid pGMI1000MP*	Bacteria	66.9	N/A	3653	137	3.6	3661	160	4.2
NC_003198	*Salmonella_typhi*	Bacteria	52.1	37C	9751	247	2.5	9756	304	3.0
NC_003197	*Salmonella_typhimurium_LT2*	Bacteria	52.2	37C	11768	348	2.9	11748	412	3.4
NC_003047	*Sinorhizobium_meliloti 1021*	Bacteria	62.7	25-30C	5896	556	8.6	5818	706	10.8
NC_003037	*Sinorhizobium_meliloti 1021 plasmid pSymA*	Bacteria	60.4	25-30C	2027	444	18.0	1905	601	24.0
NC_003078	*Sinorhizobium_meliloti 1021 plasmid pSymB*	Bacteria	62.4	25-30C	2778	407	12.8	2720	486	15.2
NC_002758	*Staphylococcus_aureus_Mu50*	Bacteria	32.9	30-37C	4718	301	6.0	4632	372	7.4
NC_002737	*Streptococcus_pyogenes_M1_GAS*	Bacteria	38.5	30-35C	3257	171	5.0	3235	188	5.5
NC_003888	*Streptomyces_coelicolor*	Bacteria	72.1	25-35C	13252	2625	16.5	9781	6199	38.8
NC_000911	*Synechocystis_PCC6803*	Bacteria	47.7	N/A	4487	210	4.5	4446	252	5.4
NC_000853	*Thermotoga_maritima*	Bacteria	46.2	80C	88	3137	97.3	59	3173	98.2
NC_004603	*Vibrio_parahaemolyticus RIMD 2210633 chromosome I*	Bacteria	45.4	20-30C	5934	112	1.9	5912	148	2.4
NC_004605	*Vibrio_parahaemolyticus RIMD 2210633 chromosome II*	Bacteria	45.4	20-30C	2882	56	1.9	2872	68	2.3
NC_003919	*Xanthomonas_citri*	Bacteria	64.8	25-30C	8813	207	2.3	8841	265	2.9

Supplementary Table 9 Predicted functions of protein products of 50 longest genes newly predicted in 7 gut metagenomic samples; 25 are from human subjects.

25 are from mice. 29 out of 50 proteins have similarity to hypothetical proteins. The entries are sorted in descending order by the length of proteins.

Index	Data source: gut community	Contig name	GC content	Gene strand	Contig length (nt)	Left end	Right end	Complete / Partial	Length in amino acids	Top hit in nr database
1	Human Subject 8	hgutS8_s8_178891	63.2	-	4310	54	1517	complete	487	gb|EBA39843.1| Hypothetical protein COLAER_00990 [Collinsella aerofaciens ATCC 25986]
2	Human Subject 7	hgutS7_s7_166367	69.2	-	1476	3	1451	partial	483	gb|EEJ20699.1| S-adenosylmethionine--tRNA ribosyltransferase-isomerase [Atopobium parvulum DSM 20469]
3	Human Subject 8	hgutS8_s8_179612	51.1	-	3453	109	1503	complete	464	gb|EEA83423.1| hypothetical protein CLONEX_00674 [Clostridium nexile DSM 1787]
4	Human Subject 7	hgutS7_s7_179341_1	56.0	-	2238	3	1277	partial	425	gb|EDM50586.1| hypothetical protein EUBVEN_01903 [Eubacterium ventriosum ATCC 27560]
5	Human Subject 7	hgutS7_s7_163800	28.9	+	2213	975	2213	partial	413	unknown protein [Candidatus Kuenenia stuttgartiensis]
6	Human Subject 7	hgutS7_s7_173828_1	57.4	+	1392	3	1196	partial	397	gb|EBA38948.1| Hypothetical protein COLAER_01809 [Collinsella aerofaciens ATCC 25986]
7	Human Subject 7	hgutS7_s7_160764	62.9	+	1611	1	1092	partial	363	gb|ABN53924.1| DNA-directed RNA polymerase, beta' subunit [Clostridium thermocellum ATCC 27405]
8	Human Subject 7	hgutS7_s7_179262	62.7	+	5770	2	1063	partial	353	dbj|BAF40243.1| hypothetical protein [Bifidobacterium adolescentis ATCC 15703]
9	Human Subject 8	hgutS8_s8_162461	41.6	+	972	3	971	partial	323	gb|EDO61545.1| hypothetical protein CLOLEP_01942 [Clostridium leptum DSM 753]
10	Human Subject 8	hgutS8_s8_161431	63.1	-	956	2	955	partial	318	dbj|BAD39344.1| ATP-dependent Clp protease ATP-binding subunit [Symbiobacterium thermophilum IAM 14863]
11	Human Subject 7	hgutS7_s7_166137	47.2	-	938	3	938	partial	312	transposase [uncultured microorganism]
12	Human Subject 7	hgutS7_s7_164528	64.7	+	7589	5396	6334	complete	312	gb|EBA39742.1| Hypothetical protein COLAER_00889 [Collinsella aerofaciens ATCC 25986]
13	Human Subject 7	hgutS7_s7_171860_1	47.3	-	916	1	915	partial	305	gb|EDR46672.1| hypothetical protein DORFOR_01895 [Dorea formicigenerans ATCC 27755]
14	Human Subject 7	hgutS7_s7_177574	55.4	+	1841	954	1841	partial	296	gb|EDR99587.1| hypothetical protein EUBSIR_02655 [Eubacterium siraeum DSM 15702]
15	Human Subject 8	hgutS8_s8_165486	46.8	-	977	3	887	partial	295	gb|EDP22564.1| hypothetical protein FAEPRAM212_00571 [Faecalibacterium prausnitzii M21/2]
16	Mouse PT6	mgutOb1_U_BO_aaa55f07_b1	59.8	+	874	3	872	partial	290	dbj|BAE74468.1| hypothetical phage protein [Sodalis glossinidius str. 'morsitans']
17	Mouse PT8	mgutLn2_U_FF_aab62a03_b1	55.7	+	870	2	868	partial	289	gb|EEK17709.1| putative penicillin-binding protein 2 [Porphyromonas uenonis 60-3]
18	Mouse PT2	mgutLn3_U_BK_aab12c04_b1	41.9	+	868	3	866	partial	288	gb|EEJ36784.1| hypothetical protein ElenDRAFT_07740 [Eggerthella lenta DSM 2243]
19	Human Subject 7	hgutS7_s7_165188	35.5	+	863	3	863	partial	287	gb|EDR47676.1| hypothetical protein DORFOR_01212 [Dorea formicigenerans ATCC 27755]
20	Mouse PT3	mgutLn1_U_BL_aaa51e08_b1	43.9	+	864	2	862	partial	287	gb|EEG52768.1| hypothetical protein CLOSTASPAR_05172 [Clostridium asparagiforme DSM 15981]
21	Human Subject 7	hgutS7_s7_179668_2	55.9	-	861	2	859	partial	286	gb|EEG93212.1| hypothetical protein ROSEINA2194_02974 [Roseburia inulinivorans DSM 16841]
22	Human Subject 7	hgutS7_s7_179256_2	67.9	-	858	1	858	partial	286	gb|EBA40043.1| Hypothetical protein COLAER_00566 [Collinsella aerofaciens ATCC 25986]
23	Mouse PT6	mgutOb1_U_BO_aaa76g07_b1	41.9	+	857	1	855	partial	285	gb|EDP18314.1| hypothetical protein CLOBOL_01382 [Clostridium bolteae ATCC BAA-613]
24	Mouse PT6	mgutOb1_U_BO_aaa54d08_b1	37.8	+	857	1	855	partial	285	gb|EEF95472.1| hypothetical protein BRYFOR_04734 [Bryantella formatexigens DSM 14469]
25	Human Subject 8	hgutS8_s8_162435	47.0	-	5399	4082	4936	complete	284	gb|EEF97952.1| hypothetical protein BRYFOR_02324 [Bryantella formatexigens DSM 14469]
26	Mouse PT3	mgutLn1_U_BL_aab11h06_b1	53.4	+	854	1	852	partial	284	gb|ACL47925.1| DNA gyrase, B subunit [Desulfovibrio desulfuricans subsp. desulfuricans str. ATCC 27774]
27	Mouse PT4	mgutOb2_U_BM_aaa80a08_b1	68.3	+	926	73	924	partial	284	gb|EEJ35868.1| o-succinylbenzoic acid synthetase [Eggerthella lenta DSM 2243]
28	Human Subject 7	hgutS7_s7_178828_1	62.3	-	2291	1440	2291	partial	283	gb|EBA38508.1| Hypothetical protein COLAER_02361 [Collinsella aerofaciens ATCC 25986]
29	Human Subject 7	hgutS7_s7_164528	64.7	+	7589	4535	5386	complete	283	gb|EEJ20894.1| fructose-bisphosphate aldolase [Atopobium parvulum DSM 20469]
30	Human Subject 8	hgutS8_s8_179668	56.0	-	3710	275	1126	complete	283	gb|EDO58706.1| hypothetical protein CLOL250_00738 [Clostridium sp. L2-50]
31	Mouse PT3	mgutLn1_U_BL_aaa18b05_b1	54.8	+	851	1	849	partial	283	gb|EDY97106.1| hypothetical protein BACPLE_00491 [Bacteroides plebeius DSM 17135]
32	Mouse PT2	mgutLn3_U_BK_aaa39f05_b1	56.2	+	850	3	848	partial	282	gb|ABR44245.1| putative phosphoribosylformylglycinamidine synthase [Parabacteroides distasonis ATCC 8503]
33	Mouse PT3	mgutLn1_U_BL_aaa31f11_b1	55.6	+	847	3	845	partial	281	gb|ACQ94146.1| hydro-lyase, Fe-S type, tartrate/fumarate subfamily, beta subunit [Tolumonas auensis DSM 9187]
34	Mouse PT3	mgutLn1_U_BL_aaa32e01_b1	41.6	+	844	3	842	partial	280	gb|EEJ53739.1| AAA+ superfamily ATPase [Mobiluncus mulieris ATCC 35243]
35	Mouse PT4	mgutOb2_U_BM_aaa28c05_b1	45.0	+	843	2	841	partial	280	gb|ABV10219.1| LPXTG cell wall surface protein [Streptococcus gordonii str. Challis substr. CH1]
36	Mouse PT6	mgutOb1_U_BO_aaa62b01_b1	49.5	+	842	1	840	partial	280	gb|EEG94524.1| hypothetical protein ROSEINA2194_01652 [Roseburia inulinivorans DSM 16841]
37	Mouse PT6	mgutOb1_U_BO_aaa36c09_b1	57.7	+	844	3	842	partial	280	gb|EEG65660.1| hypothetical protein CLOM621_04170 [Clostridium sp. M62/1]
38	Human Subject 7	hgutS7_s7_176945_1	45.6	-	839	3	839	partial	279	transposase [uncultured microorganism]
39	Mouse PT6	mgutOb1_U_BO_aaa62e02_b1	35.6	+	839	1	837	partial	279	gb|ABA89535.1| FOG: GGDEF domain protein [Pelobacter carbinolicus DSM 2380]
40	Mouse PT2	mgutLn3_U_BK_aaa82f06_b1	52.0	+	833	1	831	partial	277	gb|ACD25143.1| conserved hypothetical protein [Clostridium botulinum B str. Eklund 17B]
41	Human Subject 8	hgutS8_s8_178324	41.6	-	1080	3	827	partial	275	gb|EEP24006.1| hypothetical protein GCWU000182_03278 [Abiotrophia defectiva ATCC 49176]
42	Mouse PT2	mgutLn3_U_BK_aaa64c09_b1	59.9	+	827	1	825	partial	275	putative DNA-repair protein [Clostridiales bacterium 1_7_47_FAA]
43	Mouse PT6	mgutOb1_U_BO_aaa64h07_b1	46.6	+	828	2	826	partial	275	gb|ABI69529.1| Transposase-like protein [Syntrophomonas wolfei subsp. wolfei str. Goettingen]
44	Mouse PT8	mgutLn2_U_FF_aab09d07_b1	56.2	+	836	3	830	partial	275	gb|EDR21740.1| hypothetical protein, conserved [Entamoeba dispar SAW760]
45	Human Subject 7	hgutS7_s7_167939	61.1	-	876	50	874	partial	274	gb|EEA89649.1| hypothetical protein COLSTE_02191 [Collinsella stercoris DSM 13279]
46	Mouse PT3	mgutLn1_U_BL_aaa38h07_b1	56.4	+	824	1	822	partial	274	gb|ABR44799.1| putative N6-adenine-specific DNA methylase [Parabacteroides distasonis ATCC 8503]
47	Mouse PT4	mgutOb2_U_BM_aaa61d08_b1	59.2	+	825	2	823	partial	274	gb|ABR45053.1| carboxy-terminal processing protease precursor [Parabacteroides distasonis ATCC 8503]
48	Mouse PT6	mgutOb1_U_BO_aaa51c03_b1	45.3	+	825	2	823	partial	274	gb|EDK88996.1| DNA (cytosine-5-)-methyltransferase [Fusobacterium nucleatum subsp. polymorphum ATCC 10953]
49	Mouse PT4	mgutOb2_U_BM_aab24d06_b1	64.1	+	821	1	819	partial	273	gb|EEJ35665.1| anaerobic dehydrogenase, typically selenocysteine-containing [Eggerthella lenta DSM 2243]
50	Mouse PT6	mgutOb1_U_BO_aaa59f05_b1	38.7	+	821	1	819	partial	273	gb|EED03385.1| hypothetical protein ROSINTL182_00016 [Roseburia intestinalis L1-82]

Supplementary Table 10 The 37 genes that were used to calculate codon adaptation index.
Out of these 37 genes 23 are homologous to the *E. coli* genes used by Sharp et al. (1987). Other 14 genes were added to make up to the same number of codons as was in original set of Sharp et al. (1987). The table is sorted in descending order of the expression level. The first 31 genes are in the data set of the 100 most highly expressed genes (Supplementary Table 1). *The gene #37 homologous to an *E. coli* gene in the Sharp et al. (1987) set has a very low level of expression (0.2). It is likely to be an artifact which still should not bias computed ATS value, as in computations the codon counts from gene #37 will be added as they are.

Index	Strand	Left end	Right end	Length(nt)	Function	Expression level, log2 (Control)
1	+	119376	120563	1188	elongation factor Tu	9.9
2	-	5214543	5214833	291	30S ribosomal protein S6	9.7
3	-	3621884	3622153	270	30S ribosomal protein S15	9.3
4	+	107065	107424	360	50S ribosomal protein L7/L12	9.2
5	+	128969	129469	501	30S ribosomal protein S5	9.2
6	+	106497	106997	501	50S ribosomal protein L10	9.2
7	+	116502	116972	471	30S ribosomal protein S7	9.1
8	+	133741	134130	390	30S ribosomal protein S11	9.1
9	+	116050	116472	423	30S ribosomal protein S12	9.0
10	+	117180	119258	2079	elongation factor G	8.9
11	-	4122609	4122782	174	30S ribosomal protein S21	8.8
12	+	121305	121937	633	50S ribosomal protein L3	8.8
13	+	120962	121270	309	30S ribosomal protein S10	8.8
14	+	124437	125096	660	30S ribosomal protein S3	8.7
15	+	127583	127981	399	30S ribosomal protein S8	8.6
16	+	123796	124074	279	30S ribosomal protein S19	8.6
17	+	125743	126006	264	30S ribosomal protein S17	8.5
18	+	104969	105394	426	50S ribosomal protein L11	8.5
19	-	3642203	3642904	702	30S ribosomal protein S2	8.2
20	-	5227127	5227261	135	50S ribosomal protein L34	8.2
21	+	133351	133716	366	30S ribosomal protein S13	8.1
22	+	135291	135653	363	50S ribosomal protein L17	8.1
23	-	4082555	4082704	258	50S ribosomal protein L33	8.0
24	-	5213716	5213949	234	30S ribosomal protein S18	7.8
25	+	4459692	4460294	603	30S ribosomal protein S4	7.7
26	-	4082555	4082704	150	50S ribosomal protein L33	7.6
27	+	127368	127553	186	30S ribosomal protein S14	7.6
28	+	105572	106264	693	50S ribosomal protein L1	7.2
29	-	3641212	3642099	888	elongation factor Ts	7.1
30	-	4127654	4129489	1836	molecular chaperone DnaK	7.0
31	-	3658378	3658650	273	30S ribosomal protein S16	6.9
32	+	4136729	4136986	258	30S ribosomal protein S20	6.4
33	+	3674546	3674734	189	50S ribosomal protein L28	4.1

34	+	103778	103924	147	50S ribosomal protein L33	3.9
35	+	1436661	1437809	1149	30S ribosomal protein S1	3.8
36	-	3590395	3591753	1359	recombinase A	3.4
37	-	4153358	4153507	150	50S ribosomal protein L33*	0.2

Supplementary Table 11 The 100 most highly expressed Bacillus anthracis genes under "Control" condition as determined from RNA-Seq data.
The table is sorted in descending order of the expression level.

Index	Strand	Left end	Right end	Length(nt)	Function	Expression level, log2 (Control)
1	+	119376	120563	1188	elongation factor Tu	9.9
2	-	3299573	3299773	201	cold shock protein CspB	9.8
3	-	5214543	5214833	291	30S ribosomal protein S6	9.7
4	+	126795	127334	540	50S ribosomal protein L5	9.5
5	-	4251123	4251413	291	50S ribosomal protein L27	9.5
6	-	3621884	3622153	270	30S ribosomal protein S15	9.3
7	+	107065	107424	360	50S ribosomal protein L7/L12	9.2
8	+	128969	129469	501	30S ribosomal protein S5	9.2
9	+	106497	106997	501	50S ribosomal protein L10	9.2
10	+	116502	116972	471	30S ribosomal protein S7	9.1
11	+	896749	899193	2445	s-layer protein sap	9.1
12	+	133741	134130	390	30S ribosomal protein S11	9.1
13	+	130139	131440	1302	preprotein translocase subunit SecY	9.0
14	+	131497	132147	651	adenylate kinase	9.0
15	+	128585	128947	363	50S ribosomal protein L18	9.0
16	+	116050	116472	423	30S ribosomal protein S12	9.0
17	+	129483	129665	183	50S ribosomal protein L30	9.0
18	+	117180	119258	2079	elongation factor G	8.9
19	-	1606994	1607857	864	flagellin	8.9
20	+	132962	133180	219	translation initiation factor IF-1	8.8
21	-	4122609	4122782	174	30S ribosomal protein S21	8.8
22	+	121305	121937	633	50S ribosomal protein L3	8.8
23	+	120962	121270	309	30S ribosomal protein S10	8.8
24	+	122905	123735	831	50S ribosomal protein L2	8.8
25	+	126457	126768	312	50S ribosomal protein L24	8.8
26	+	134311	135255	945	DNA-directed RNA polymerase subunit alpha	8.7
27	+	124437	125096	660	30S ribosomal protein S3	8.7
28	+	127583	127981	399	30S ribosomal protein S8	8.6
29	+	123796	124074	279	30S ribosomal protein S19	8.6
30	+	124092	124433	342	50S ribosomal protein L22	8.6

31	+	125743	126006	264	30S ribosomal protein S17	8.5
32	+	104969	105394	426	50S ribosomal protein L11	8.5
33	-	4122150	4122593	444	gatb/yqey domain-containing protein	8.5
34	+	122586	122876	291	50S ribosomal protein L23	8.5
35	+	125098	125532	435	50S ribosomal protein L16	8.5
36	-	4863474	4864478	1005	glyceraldehyde-3-phosphate dehydrogenase	8.4
37	-	4636892	4637092	201	cold shock protein CspD	8.4
38	+	133216	133329	114	50S ribosomal protein L36	8.4
39	+	126050	126418	369	50S ribosomal protein L14	8.3
40	+	121963	122586	624	50S ribosomal protein L4	8.3
41	-	3666095	3666328	234	acyl carrier protein	8.2
42	-	3642203	3642904	702	30S ribosomal protein S2	8.2
43	-	5227127	5227261	135	50S ribosomal protein L34	8.2
44	+	128014	128553	540	50S ribosomal protein L6	8.2
45	+	133351	133716	366	30S ribosomal protein S13	8.1
46	-	4251773	4252081	309	50S ribosomal protein L21	8.1
47	+	135291	135653	363	50S ribosomal protein L17	8.1
48	+	129699	130139	441	50S ribosomal protein L15	8.0
49	-	3744933	3745106	174	50S ribosomal protein L32	7.9
50	-	5213995	5214516	522	single-stranded DNA-binding protein	7.9
51	-	4648464	4649408	945	L-lactate dehydrogenase	7.8
52	-	5058593	5058838	246	50S ribosomal protein L31 type B	7.8
53	-	5213716	5213949	234	30S ribosomal protein S18	7.8
54	+	108391	111924	3534	DNA-directed RNA polymerase subunit beta	7.7
55	+	125522	125722	201	50S ribosomal protein L29	7.7
56	+	132147	132893	747	methionine aminopeptidase	7.7
57	+	258543	258827	285	co-chaperonin GroES	7.7
58	+	4459692	4460294	603	30S ribosomal protein S4	7.7
59	-	1544452	1544649	198	cold shock protein CspB	7.6
60	-	4082555	4082704	150	50S ribosomal protein L33	7.6
61	+	127368	127553	186	30S ribosomal protein S14	7.6
62	+	1447230	1447502	273	DNA-binding protein HU	7.5
63	-	5037510	5037911	402	F0F1 ATP synthase subunit epsilon	7.5
64	+	111962	115573	3612	DNA-directed RNA polymerase subunit beta'	7.5
65	+	3530610	3530882	273	DNA-binding protein HU	7.5
66	+	226255	226362	108	hypothetical protein	7.4
67	-	4911264	4911806	543	ribosomal subunit interface protein	7.4
68	+	2047217	2047489	273	hypothetical protein	7.4
69	-	4858510	4859805	1296	phosphopyruvate hydratase	7.3
70	+	258866	260500	1635	chaperonin GroEL	7.3
71	-	4380540	4380740	201	50S ribosomal protein L35	7.3
72	-	4862150	4863334	1185	phosphoglycerate kinase	7.3
73	-	4861362	4862117	756	triosephosphate isomerase	7.2

74	-	4859836	4861365	1530	phosphoglyceromutase	7.2
75	+	105572	106264	693	50S ribosomal protein L1	7.2
76	-	3830293	3831270	978	pyruvate dehydrogenase complex E1 component, beta subunit	7.2
77	-	3658136	3658363	228	kh domain-containing protein	7.2
78	+	139149	139586	438	50S ribosomal protein L13	7.2
79	-	3641212	3642099	888	elongation factor Ts	7.1
80	-	4251417	4251707	291	hypothetical protein	7.1
81	+	503601	505850	2250	formate acetyltransferase	7.1
82	+	505920	506651	732	pyruvate formate-lyase-activating enzyme	7.1
83	+	1838333	1839595	1263	NLP/P60 family protein	7.1
84	-	4093878	4094489	612	superoxide dismutase, Mn	7.0
85	-	4127654	4129489	1836	molecular chaperone DnaK	7.0
86	-	3831274	3832389	1116	pyruvate dehydrogenase complex E1 component, alpha subunit	7.0
87	-	3656275	3656619	345	50S ribosomal protein L19	6.9
88	-	3658378	3658650	273	30S ribosomal protein S16	6.9
89	+	226405	226515	111	hypothetical protein	6.9
90	-	3828941	3830200	1260	branched-chain alpha-keto acid dehydrogenase subunit E2	6.9
91	-	3901931	3903643	1713	phosphoenolpyruvate-protein phosphotransferase	6.9
92	-	5037932	5039341	1410	F0F1 ATP synthase subunit beta	6.9
93	-	5063560	5064417	858	fructose-bisphosphate aldolase	6.9
94	+	120795	120962	168	hypothetical protein	6.8
95	-	2690701	2690793	93	hypothetical protein	6.8
96	-	722230	723105	876	quinol oxidase, subunit II	6.7
97	-	4021331	4021723	393	hypothetical protein	6.7
98	-	355897	356460	564	alkyl hydroperoxide reductase subunit C	6.7
99	-	3827523	3828935	1413	dihydrolipoamide dehydrogenase	6.7
100	-	5040657	5042165	1509	F0F1 ATP synthase subunit alpha	6.7

REFERENCES

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-3402.

Antonov, I. and M. Borodovsky (2010). "GeneTack: Frameshift identification in protein coding sequences by the Viterbi algorithm." Journal of Bioinformatics and Computational Biology **8**(3): 1-17.

Audic, S. and J. M. Claverie (1998). "Self-identification of protein-coding regions in microbial genomes." Proc Natl Acad Sci U S A **95**(17): 10026-10031.

Azad, R. K. and M. Borodovsky (2004). "Effects of choice of DNA sequence model structure on gene identification accuracy." Bioinformatics **20**(7): 993-1005.

Badger, J. H. and G. J. Olsen (1999). "CRITICA: coding region identification tool invoking comparative analysis." Mol Biol Evol **16**(4): 512-524.

Bairoch, A. and R. Apweiler (2000). "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000." Nucleic Acids Res **28**(1): 45-48.

Bakkeren, G., W. Zhu, et al. (2010). "Gene prediction in Puccinia triticina based on EST data." **In Preparation**.

Basak, S., T. Banerjee, et al. (2004). "Investigation on the causes of codon and amino acid usages variation between thermophilic Aquifex aeolicus and mesophilic Bacillus subtilis." J Biomol Struct Dyn **22**(2): 205-214.

Benson, G. (1999). "Tandem repeats finder: a program to analyze DNA sequences." Nucleic Acids Res **27**(2): 573-580.

Bernaola-Galvan, P., J. L. Oliver, et al. (2004). "Quantifying intrachromosomal GC heterogeneity in prokaryotic genomes." Gene **333**: 121-133.

Besemer, J. and M. Borodovsky (1999). "Heuristic approach to deriving models for gene finding." Nucleic Acids Res **27**(19): 3911-3920.

Besemer, J., A. Lomsadze, et al. (2001). "GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions." Nucleic Acids Res **29**(12): 2607-2618.

Borodovsky, M., E. V. Koonin, et al. (1994). "New genes in old sequence: a strategy for finding genes in the bacterial genome." Trends Biochem Sci **19**(8): 309-313.

Borodovsky, M. and J. Mcininch (1993). "Genmark - Parallel gene recognition for both DNA strands." Computers & Chemistry **17**(2): 123-133.

Borodovsky, M., Y. A. Sprizhitskii, et al. (1986). "Statistical Patterns in Primary Structures of the Functional Regions of the Genome in Escherichia Coli." Molecular Biology **20**: 826-833, 833-840, 1144-1150.

Bove, J. M. (1993). "Molecular features of mollicutes." Clin Infect Dis **17 Suppl 1**: S10-31.

Brown, N. P., C. Sander, et al. (1998). "Frame: detection of genomic sequencing errors." Bioinformatics **14**(4): 367-371.

Chen, K. and L. Pachter (2005). "Bioinformatics for whole-genome shotgun sequencing of microbial communities." PLoS Comput Biol **1**(2): 106-112.

Chen, S. L., W. Lee, et al. (2004). "Codon usage between genomes is constrained by genome-wide mutational processes." Proceedings of the National Academy of Sciences of the United States of America **101**(10): 3480-3485.

Claverie, J. M., O. Poirot, et al. (1997). "The difficulty of identifying genes in anonymous vertebrate sequences." Comput Chem **21**(4): 203-214.

Cole, J. R., Q. Wang, et al. (2009). "The Ribosomal Database Project: improved alignments and new tools for rRNA analysis." Nucleic Acids Res **37**(Database issue): D141-145.

Cole, S. T., K. Eiglmeier, et al. (2001). "Massive gene decay in the leprosy bacillus." Nature **409**(6823): 1007-1011.

Crick, F. H. (1966). "Codon--anticodon pairing: the wobble hypothesis." J Mol Biol **19**(2): 548-555.

Cristianini, N. and J. Shawe-Taylor (2000). An introduction to support Vector Machines: and other kernel-based learning methods, Cambridge Univ Pr.

Delcher, A. L., K. A. Bratke, et al. (2007). "Identifying bacterial genes and endosymbiont DNA with Glimmer." Bioinformatics **23**(6): 673-679.

Delcher, A. L., D. Harmon, et al. (1999). "Improved microbial gene identification with GLIMMER." Nucleic Acids Res **27**(23): 4636-4641.

Fickett, J. W. (1982). "Recognition of protein coding regions in DNA sequences." Nucleic Acids Res **10**(17): 5303-5318.

Fickett, J. W. and C. S. Tung (1992). "Assessment of protein coding measures." Nucleic Acids Res **20**(24): 6441-6450.

Finn, R. D., J. Tate, et al. (2008). "The Pfam protein families database." Nucleic Acids Res **36**(Database issue): D281-288.

Fleischmann, R. D., M. D. Adams, et al. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." Science **269**(5223): 496-512.

Frishman, D., A. Mironov, et al. (1998). "Combining diverse evidence for gene recognition in completely sequenced bacterial genomes." Nucleic Acids Res **26**(12): 2941-2947.

Gill, S. R., M. Pop, et al. (2006). "Metagenomic analysis of the human distal gut microbiome." Science **312**(5778): 1355-1359.

Gorban, A. N. and A. Y. Zinovyev (2007). "The mystery of two straight lines in bacterial genome statistics." Bull Math Biol **69**(7): 2429-2442.

Gribskov, M., J. Devereux, et al. (1984). "The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression." Nucleic Acids Res **12**(1 Pt 2): 539-549.

Hayes, W. S. and M. Borodovsky (1998). "Deriving ribosomal binding site (RBS) statistical models from unannotated DNA sequences and the use of the RBS model for N-terminal prediction." Pac Symp Biocomput: 279-290.

Hayes, W. S. and M. Borodovsky (1998). "How to interpret an anonymous bacterial genome: machine learning approach to gene identification." Genome Res **8**(11): 1154-1171.

Hoff, K. J. (2009). "The effect of sequencing errors on metagenomic gene prediction." BMC Genomics **10**: 520.

Hoff, K. J., T. Lingner, et al. (2009). "Orphelia: predicting genes in metagenomic sequencing reads." Nucleic Acids Res **37**(Web Server issue): W101-105.

Hoff, K. J., M. Tech, et al. (2008). "Gene prediction in metagenomic fragments: a large scale machine learning approach." BMC Bioinformatics **9**: 217.

Hu, G. Q., J. T. Guo, et al. (2009). "MetaTISA: Metagenomic Translation Initiation Site Annotator for improving gene start prediction." Bioinformatics **25**(14): 1843-1845.

Hu, G. Q., X. Zheng, et al. (2008). "Computational evaluation of TIS annotation for prokaryotic genomes." BMC Bioinformatics **9**: 160.

Hu, G. Q., X. Zheng, et al. (2008). "ProTISA: a comprehensive resource for translation initiation site annotation in prokaryotic genomes." Nucleic Acids Res **36**(Database issue): D114-119.

Hu, G. Q., X. Zheng, et al. (2009). "Prediction of translation initiation site for microbial genomes with TriTISA." Bioinformatics **25**(1): 123-125.

Jansen, R., H. J. Bussemaker, et al. (2003). "Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models." Nucleic Acids Res **31**(8): 2242-2251.

Jukes, T. H. and S. Osawa (1993). "Evolutionary changes in the genetic code." Comp Biochem Physiol B **106**(3): 489-494.

Kanaya, S., Y. Yamada, et al. (1999). "Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis." Gene **238**(1): 143-155.

Karro, J. E., Y. Yan, et al. (2007). "Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation." Nucleic Acids Res **35**(Database issue): D55-60.

Kato, M., K. Nishikawa, et al. (1990). "The difference in the type of codon-anticodon base pairing at the ribosomal P-site is one of the determinants of the translational rate." J Biochem **107**(2): 242-247.

Kattenhorn, L. M., R. Mills, et al. (2004). "Identification of proteins associated with murine cytomegalovirus virions." Journal of Virology **78**(20): 11187-11197.

Kislyuk, A., A. Lomsadze, et al. (2009). "Frameshift detection in prokaryotic genomic sequences." Int J Bioinform Res Appl **5**(4): 458-477.

Knight, R. D., S. J. Freeland, et al. (2001). "A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes." Genome Biol **2**(4): research0010.0011-0010.0013.

Krause, L., N. N. Diaz, et al. (2006). "Finding novel genes in bacterial communities isolated from the environment." Bioinformatics **22**(14): e281-289.

Krause, L., A. C. McHardy, et al. (2007). "GISMO--gene identification using a support vector machine for ORF classification." Nucleic Acids Res **35**(2): 540-549.

Krogh, A. (1997). "Two methods for improving performance of an HMM and their application for gene finding." Proc Int Conf Intell Syst Mol Biol **5**: 179-186.

Krogh, A. (2000). "Using database matches with for HMMGene for automated gene detection in Drosophila." Genome Res **10**(4): 523-528.

Kullback, S. and R. Leibler (1951). "On information and sufficiency." The Annals of Mathematical Statistics: 79-86.

Lam, H. Y., E. Khurana, et al. (2009). "Pseudofam: the pseudogene families database." Nucleic Acids Res **37**(Database issue): D738-743.

Larsen, T. S. and A. Krogh (2003). "EasyGene - a prokaryotic gene finder that ranks ORFs by statistical significance." Bmc Bioinformatics **4**: 15.

Lawrence, C. E., S. F. Altschul, et al. (1993). "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment." Science **262**(5131): 208-214.

Liolios, K., I. M. Chen, et al. (2010). "The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata." Nucleic Acids Res **38**(Database issue): D346-354.

Lithwick, G. and H. Margalit (2003). "Hierarchy of sequence-dependent features associated with prokaryotic translation." Genome Res **13**(12): 2665-2673.

Liu, J. S. and C. E. Lawrence (1999). "Bayesian inference on biopolymer models." Bioinformatics **15**(1): 38-52.

Lobry, J. R. and A. Necsulea (2006). "Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes." Gene **385**: 128-136.

Lomsadze, A., V. Ter-Hovhannisyan, et al. (2005). "Gene identification in novel eukaryotic genomes by self-training algorithm." Nucleic Acids Res **33**(20): 6494-6506.

Lowe, T. M. and S. R. Eddy (1997). "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence." Nucleic Acids Res **25**(5): 955-964.

Lukashin, A. V. and M. Borodovsky (1998). "GeneMark.hmm: new solutions for gene finding." Nucleic Acids Res **26**(4): 1107-1115.

Mahony, S., J. O. McInerney, et al. (2004). "Gene prediction using the Self-Organizing Map: automatic generation of multiple gene models." BMC Bioinformatics **5**: 23.

Mardis, E. R. (2008). "The impact of next-generation sequencing technology on genetics." Trends Genet **24**(3): 133-141.

Markowitz, V. M., N. N. Ivanova, et al. (2008). "IMG/M: a data management and analysis system for metagenomes." Nucleic Acids Res **36**(Database issue): D534-538.

Martin, J., W. Zhu, et al. (2009). "Assessment of Gene Annotation Accuracy by Inferring Transcripts from RNA-Seq." BIBM 2009: 54-59.

Martin, J., W. Zhu, et al. (2010). "Bacillus anthracis genome organization in light of whole transcriptome sequencing." BMC Bioinformatics **11**(Suppl 3): S10.

Mavromatis, K., N. Ivanova, et al. (2007). "Use of simulated data sets to evaluate the fidelity of metagenomic processing methods." Nat Methods **4**(6): 495-500.

Mills, R., M. Rozanov, et al. (2003). "Improving gene annotation of complete viral genomes." Nucleic Acids Res **31**(23): 7041-7055.

Nakabachi, A., A. Yamashita, et al. (2006). "The 160-kilobase genome of the bacterial endosymbiont Carsonella." Science **314**(5797): 267.

Nekrutenko, A. and W. H. Li (2000). "Assessment of compositional heterogeneity within and between eukaryotic genomes." Genome Res **10**(12): 1986-1995.

Nelson, K. E., R. A. Clayton, et al. (1999). "Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of Thermotoga maritima." Nature **399**(6734): 323-329.

Neuwald, A. F., J. S. Liu, et al. (1995). "Gibbs motif sampling: detection of bacterial outer membrane protein repeats." Protein Sci **4**(8): 1618-1632.

Newberg, L. A., W. A. Thompson, et al. (2007). "A phylogenetic Gibbs sampler that yields centroid solutions for cis-regulatory site prediction." Bioinformatics **23**(14): 1718-1727.

Nielsen, P. and A. Krogh (2005). "Large-scale prokaryotic gene prediction and comparison to genome annotation." Bioinformatics **21**(24): 4322-4329.

Noguchi, H., J. Park, et al. (2006). "MetaGene: prokaryotic gene finding from environmental genome shotgun sequences." Nucleic Acids Research **34**(19): 5623-5630.

Noguchi, H., T. Taniguchi, et al. (2008). "MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes." DNA Res **15**(6): 387-396.

Pertea, M., K. Ayanbule, et al. (2009). "OperonDB: a comprehensive database of predicted operons in microbial genomes." Nucleic Acids Res **37**(Database issue): D479-482.

Posfai, J. and R. J. Roberts (1992). "Finding errors in DNA sequences." Proc Natl Acad Sci U S A **89**(10): 4698-4702.

Pruesse, E., C. Quast, et al. (2007). "SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB." Nucleic Acids Res **35**(21): 7188-7196.

Rabiner, L. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition." Proceedings of the IEEE **77**(2): 257-286.

Randau, L., R. Munch, et al. (2005). "Nanoarchaeum equitans creates functional tRNAs from separate genes for their 5'- and 3'-halves." Nature **433**(7025): 537-541.

Reese, M. G., D. Kulp, et al. (2000). "Genie--gene finding in Drosophila melanogaster." Genome Res **10**(4): 529-538.

Rocha, E. P. (2004). "Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization." Genome Res **14**(11): 2279-2286.

Rohl, C. A., W. Fiori, et al. (1999). "Alanine is helix-stabilizing in both template-nucleated and standard peptide helices." Proc Natl Acad Sci U S A **96**(7): 3682-3687.

Rudd, K. E. (2000). "EcoGene: a genome sequence database for Escherichia coli K-12." Nucleic Acids Res **28**(1): 60-64.

Rudd, K. E., W. Miller, et al. (1991). "Mapping sequenced E.coli genes by computer: software, strategies and examples." Nucleic Acids Res **19**(3): 637-647.

Rudner, R., J. Karkas, et al. (1968). "Separation of B. subtilis DNA into complementary strands, III. Direct analysis." Proceedings of the National Academy of Sciences **60**(3): 921-922.

Rudner, R., J. D. Karkas, et al. (1968). "Separation of B. subtilis DNA into complementary strands. 3. Direct analysis." Proc Natl Acad Sci U S A **60**(3): 921-922.

Salzberg, S. L., A. L. Delcher, et al. (1998). "Microbial gene identification using interpolated Markov models." Nucleic Acids Res **26**(2): 544-548.

Sayers, E. W., T. Barrett, et al. (2010). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res **38**(Database issue): D5-16.

Sayers, E. W., T. Barrett, et al. (2009). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res **37**(Database issue): D5-15.

Schiex, T., J. Gouzy, et al. (2003). "FrameD: A flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences." Nucleic Acids Res **31**(13): 3738-3741.

Sharp, P. M. and W. H. Li (1987). "The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications." Nucleic Acids Res **15**(3): 1281-1295.

Shine, J. and L. Dalgarno (1974). "The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites." Proc Natl Acad Sci U S A **71**(4): 1342-1346.

Shmatkov, A. M., A. A. Melikyan, et al. (1999). "Finding prokaryotic genes by the 'frame-by-frame' algorithm: targeting gene starts and overlapping genes." Bioinformatics **15**(11): 874-886.

Skovgaard, M., L. J. Jensen, et al. (2001). "On the total number of genes and their length distribution in complete microbial genomes." Trends Genet **17**(8): 425-428.

Slupska, M. M., A. G. King, et al. (2001). "Leaderless transcripts of the crenarchaeal hyperthermophile Pyrobaculum aerophilum." J Mol Biol **309**(2): 347-360.

Staden, R. (1984). "Computer Methods to Locate Signals in Nucleic-Acid Sequences." Nucleic Acids Research **12**(1): 505-519.

Staden, R. (1984). "Measurements of the Effects That Coding for a Protein Has on a DNA-Sequence and Their Use for Finding Genes." Nucleic Acids Research **12**(1): 551-567.

Stein, L. D., C. Mungall, et al. (2002). "The generic genome browser: a building block for a model organism system database." Genome Res **12**(10): 1599-1610.

Steitz, J. A., Jakes, K. (1975). "How ribosomes select initiator regions in mRNA: base pair formation between the 3' terminus of 16S rRNA and the mRNA during initiation of protein synthesis in *Escherichia coli*." Proc. Natl. Acad. Sci. **72**: 4734-4738.

Sueoka, N. (1962). "On the genetic basis of variation and heterogeneity of DNA base composition." Proc Natl Acad Sci U S A **48**: 582-592.

Suzek, B. E., M. D. Ermolaeva, et al. (2001). "A probabilistic method for identifying start codons in bacterial genomes." Bioinformatics **17**(12): 1123-1130.

Tech, M. and R. Merkl (2003). "YACOP: Enhanced gene prediction obtained by a combination of existing methods." In Silico Biol **3**(4): 441-451.

Tech, M., N. Pfeifer, et al. (2005). "TICO: a tool for improving predictions of prokaryotic translation initiation sites." Bioinformatics **21**(17): 3568-3569.

Ter-Hovhannisyan, V., A. Lomsadze, et al. (2008). "Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training." Genome Res **18**(12): 1979-1990.

Thompson, W., M. J. Palumbo, et al. (2004). "Decoding human regulatory circuits." Genome Res **14**(10A): 1967-1974.

Thompson, W., E. C. Rouchka, et al. (2003). "Gibbs Recursive Sampler: finding transcription factor binding sites." Nucleic Acids Res **31**(13): 3580-3585.

Turnbaugh, P. J. (2006). "An obesity-associated gut microbiome with increased capacity for energy harvest." Nature **444**: 1027-1031.

Venter, J. C., K. Remington, et al. (2004). "Environmental genome shotgun sequencing of the Sargasso Sea." Science **304**(5667): 66-74.

Xu, Y., R. Mural, et al. (1994). "Recognizing exons in genomic sequence using GRAIL II." Genet Eng (N Y) **16**: 241-253.

Xu, Y. and E. C. Uberbacher (1996). "Gene prediction by pattern recognition and homology search." Proc Int Conf Intell Syst Mol Biol **4**: 241-251.

Yada, T. and M. Hirosawa (1996). "Detection of short protein coding regions within the cyanobacterium genome: application of the hidden Markov model." DNA Res **3**(6): 355-361.

Yada, T., Y. Totoki, et al. (2001). "A novel bacterial gene-finding system with improved accuracy in locating start codons." DNA Research **8**(3): 97-106.

Yooseph, S., W. Li, et al. (2008). "Gene identification and protein classification in microbial metagenomic sequence data via incremental clustering." BMC Bioinformatics **9**: 182.

Yooseph, S., G. Sutton, et al. (2007). "The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families." PLoS Biol **5**(3): e16.

Zavala, A., H. Naya, et al. (2002). "Trends in codon and amino acid usage in Thermotoga maritima." J Mol Evol **54**(5): 563-568.

Zhou, J., W. J. Liu, et al. (1999). "Papillomavirus capsid protein expression level depends on the match between codon usage and tRNA availability." J Virol **73**(6): 4972-4982.

Zhu, H., G. Q. Hu, et al. (2007). "MED: a new non-supervised gene prediction algorithm for bacterial and archaeal genomes." BMC Bioinformatics **8**: 97.

Zhu, W., A. Lomsadze, et al. (2010). "ab initio Gene Identification in Metagenomic Sequences." Nucl. Acids Res. **In Press**.

Zhu, W., A. Lomsadze, et al. (2010). "GeneMarkS Plus: Improving gene annotation in complete prokaryotic genomes." **In Preparation**.

This Page Intentionally Left Blank